DATA GOVERNANCE GUIDEBOOK

IN-4-AHA Project - *Innovation Networks for Scaling Active and Healthy Ageing*

Work Package: WP5
Deliverable: 5.3
Dissemination level: Public
Version: 3.0, 25.06.2022

2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017603

http://ec.europa.eu/digital-single-market/ehealth

Innovation Networks for Active and Healthy Ageing (IN-4-AHA) is a project funded by the European Commission under the Horizon 2020 programme Coordination and Support Action (CSA), Grant Agreement No. 101017603.

This document has been prepared within work package 5 (*WP5*) task 5.3 of the project.

More information about the project can be found on the IN-4-AHA webpage and social media pages:
https://innovation4ageing.eu/
https://www.facebook.com/IN4AHA
https://twitter.com/EIP_AHA
https://www.linkedin.com/groups/8912125/

More information about the EIP on AHA community and FUTURIUM platform:
https://futurium.ec.europa.eu/en/active-and-healthy-living-digital-world
https://digital-strategy.ec.europa.eu/en/policies/eip-aha

Disclaimer
The document reflects only the authors´ view, and the European Commission is not responsible for any use that may be made of the information it contains.

Authors
Andres Kütt, Kaspar Kala, Florian Marcus, Hille Hinsberg (Proud Engineers)

Revised and contributed by

Organisation	Name
Sorainen Law Office	Lise-Lotte Lääne
Project partners	Comments submitted to pre-final draft in June 2022

History of changes

Version	Date	Modifications
1.0	28.04.2022	
2.0	27.05.2022	Revisions and additions
3.0	20.6.2022	Formatting, revisions according to comments

Data Governance Guidebook

Glossary

- **Data** is digitally stored statements about the world.
- **Data item** (also, data element) is a collection of data facts constituting a meaningful business record and confirming to a particular semantic definition.
- **Data lifecycle** is a process by which data items travel through the organisation.
- **Dataset** is a collection of data items.
- **Data management** is a term used to describe how organisations manage and influence the collection and utilisation of data. Data management is the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles (DAMA International, 2009).
- **Data governance** is the exercise of authority and control over the management of data (DAMA International, 2009).
- **Data subject** means identified or identifiable natural person[s] (GDPR). In other words, people from whom or about whom you collect information in connection with providing services and other business operations.
- **Data controller** is a natural or legal person, public authority, agency, or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data (GDPR).
- **Data processor** is a natural or legal person, public authority, agency, or other body which processes personal data on behalf of the controller (GDPR).
- **Data sharing** is used as a generic term by which parties other than the original controller can process the data of that controller (European Commission, 2022).
- **Consent of the data subject** means any freely given, specific, informed, and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her. (GDPR: European Commission, 2016). Consent must be freely given and fully informed, this means the purpose of processing data and all types of processing planned must be made clear to the data subject in a concise, user-friendly, and easily understandable way at the time at which data are collected.
- **General Data Protection Regulation (GDPR)** is a regulation in EU law strengthening and harmonising EU/EEA procedures concerning the collection, storage, processing, access, use, transfer, and erasure of personal data.
- **Health data:** GDPR Article 4(15) defines data concerning health as personal data related to the physical and mental health of a natural person, including the provision of health care services, which reveal information about his or her health status. The data generated in the context of healthcare includes both personal data as defined in Article 4(1) GDPR, and sensitive personal data as defined in Article 9(1) GDPR. In

practice, health data are often understood as any personal data generated within healthcare systems, including data concerning health which are collected through wearable devices, apps, and self-reported information. In this guidebook, a wide definition of health data is used to include all the above, as well genetic data and biometric data.

- **Healthcare:** Health and social care are understood in the sense of article 9(2)(h) GDPR, to include direct care provision, also long-term care (EC, 2021). For the sake of simplicity, the term 'healthcare' is used to include all types of patient care, including medical or social care (European Commission, 2021).
- **Healthcare provider** is defined as any natural or legal person or any other entity legally providing healthcare on the territory of a Member State, in accordance with Directive 2011/24/EU.
- **Personal data** means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. (GDPR, Article 4(1))
- **Data processing:** any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction (GDPR).
- **Data quality:** comprehensive view of usefulness of data to support decision making. Data quality is defined as "fitness for use" for users' needs. The OECD views quality in terms of dimensions such as relevance, accuracy, credibility, timeliness, accessibility, interpretability and coherence and cost-efficiency (HIMSS Dictionary of Healthcare Information Technology, OECD, 2015)
- **Privacy** is freedom of intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual (ISO TS/82304–2, 2021).
- **Metadata** is data about data an organisation has, e.g., what it represents, how it is classified, where it came from, how it moves within the organisation, how it evolves through use, who can and cannot use it, and whether it is of high quality, etc. (DAMA International, 2009). Metadata is also data and is subject to the same data governance model on a different abstraction level.
- **Data lineage:** a pathway along which it moves from its point of origin to its point of usage, sometimes called the data chain (DAMA International, 2009).
- **European Health Data Space (EHDS)**: Proposal for a regulation on the European Health Data Space, addressing health-specific challenges to electronic health data access and sharing (European Commission, 2022)

- **Data Management Body of Knowledge (DMBOK):** recommended as a thorough handbook, by DAMA International (DAMA International, 2009).

1. Introduction

1.1. About the IN4AHA project

The project Innovation Networks for Active and Healthy Ageing (IN-4-AHA) aims to help bring tested and ready-to-use applications from one country to cross-border use and strengthen the supporting role and operating model of health technology clusters in the field of active and healthy ageing (AHA). The project's supporting measures are tailored to the ecosystem of innovators in the AHA domain in Europe. The focus is on the design and implementation support for scaling-up innovative solutions that are tested in a specific region or state and are ready to expand to a larger market.

The innovation deployment supporting and enabling role is taken by health technology clusters that perform a variety of governance functions in their network of innovation actors – including the service providers (SP), the local, regional, and national authorities, the funders and the regulators, and the communities of practice through Reference Sites connected with European Innovation Partnership in the AHA domain. Throughout the project special attention is given to the challenges of getting an innovative product into active service and scale up in the market.

1.2. Why are the guidelines needed?

Healthcare relies on digital technologies to support care delivery (at a service organisation level as well as throughout the healthcare system). According to WHO guidance on scale-up strategy, successful scaling cannot happen without locally generated evidence on effectiveness and feasibility of innovative solutions (WHO, 2009). Good data governance can deliver practical insights and support scaling strategies of service providers. However, most organisations find it challenging to manage the use and exchange of data and organise available data assets. These guidelines aim to improve data management practice in the ecosystem of the active and healthy ageing.

This ecosystem comprises four tiers with dedicated functions and roles within the ecosystem:

a) the first tier includes Service Providers who are directly involved in the service delivery for end users and therefore carry the responsibility of being the **data stewards** in relation with customer data. Service Providers do not own the data, but are the custodians of the data assets, ensuring the quality, accuracy and security of the data collected in service processes.

b) the second tier are **data facilitators** (data intermediaries) who provide data mediation services, i.e., create new value for the ecosystem by **capturing aspects of existing data sets** and deploying appropriate methods for re-using them. Intermediation services may include platforms or databases enabling the exchange or joint exploitation of data and the establishment of specific infrastructure for the interconnection of data holders and data users. The Data Governance Act creates enablers for improved availability of data by trusted data intermediaries and by strengthening data-sharing mechanisms across the EU (European Commission, 2020).

> *EXAMPLE*
> *Findata.fi (Finland) facilitates the secondary use of health and social data, including for the purpose of development and innovation operations. Findata makes data permit and data request decisions regarding the data of other controllers. It takes the responsibility for the gathering, combining, previewing and disclosing of data for secondary use.*
> *Health Data Hub (France) supports data re-users in technical and regulatory procedures and provides an access to a catalogue of databases associated with reimbursement from health insurance. The Hub staff works with national ethics and scientific committee (CESREES) which assesses the public interest of the purpose pursued, relevance and adequacy of the data requested with this purpose, the foreseen methodology for data use. They also develop cataloging, metadata, documentation and synthetic data creation tools in collaboration with data managers.*

c) the third-tier actors are regional and national **policymakers** that make the rules for health and care providers on how to manage their data assets, including **enforcing authentication and access rights to data** and ensuring compliance with laws and regulations.

d) the fourth-tier actors are **regulators** on the EU level who **set forth regulatory and architecture frameworks** to enable exchange of data and services across border within the EU.

This guide is designed first and foremost to support the first tier, the innovators, the providers of services in the AHA domain in deploying digital and data-driven technology for use in health and care, so that data management can be built into the business model and service development 'by design'. Its main goal is to support the value creation for the users and adhere to key principles of privacy, security, interoperability, and effectiveness.

1.3. Scope of the guidebook

To describe the scope of the document, basic definitions should be provided first. Conversely, defining data, our main subject, is quite difficult with no single dominant definition having emerged (See Dama International, 2019 for discussion and references). In this guidebook, we define data as digitally stored statements about the world. This definition combines the role of data in representing facts about the world with the notion that not all data is factually correct, while limiting the scope of the definition to only things that can be processed digitally.

In itself, data is not very useful: the weight of a person or the number of sparrows in Peru alone does not provide much value. Data can be seen as raw material requiring meaning to become information (Silver and Silver, 1973) i.e., answers to specific questions. For example, the data points referenced above can be used to answer questions about whether somebody is overweight and if the Peruvian sparrow population is declining or not. From information, knowledge (e.g., knowledge on the obesity epidemic) can be synthesised that can lead to wisdom. In this context, our document focuses only on data.

Most human endeavours can be seen as a combination of value creation and value capture, doing something useful and being able to reap the benefits. Be it financial income or the joy of seeing your child learn to ride the bicycle, the benefit leads us to do more of the useful thing we were doing.

This guidebook focuses on value creation through data governance rather than capturing its value. This choice is made deliberately because, as an internal process to an organisation, **data governance can be supported by common best practices and recommendations whereas value capture as part of a particular business model is intrinsically unique for every organisation.**

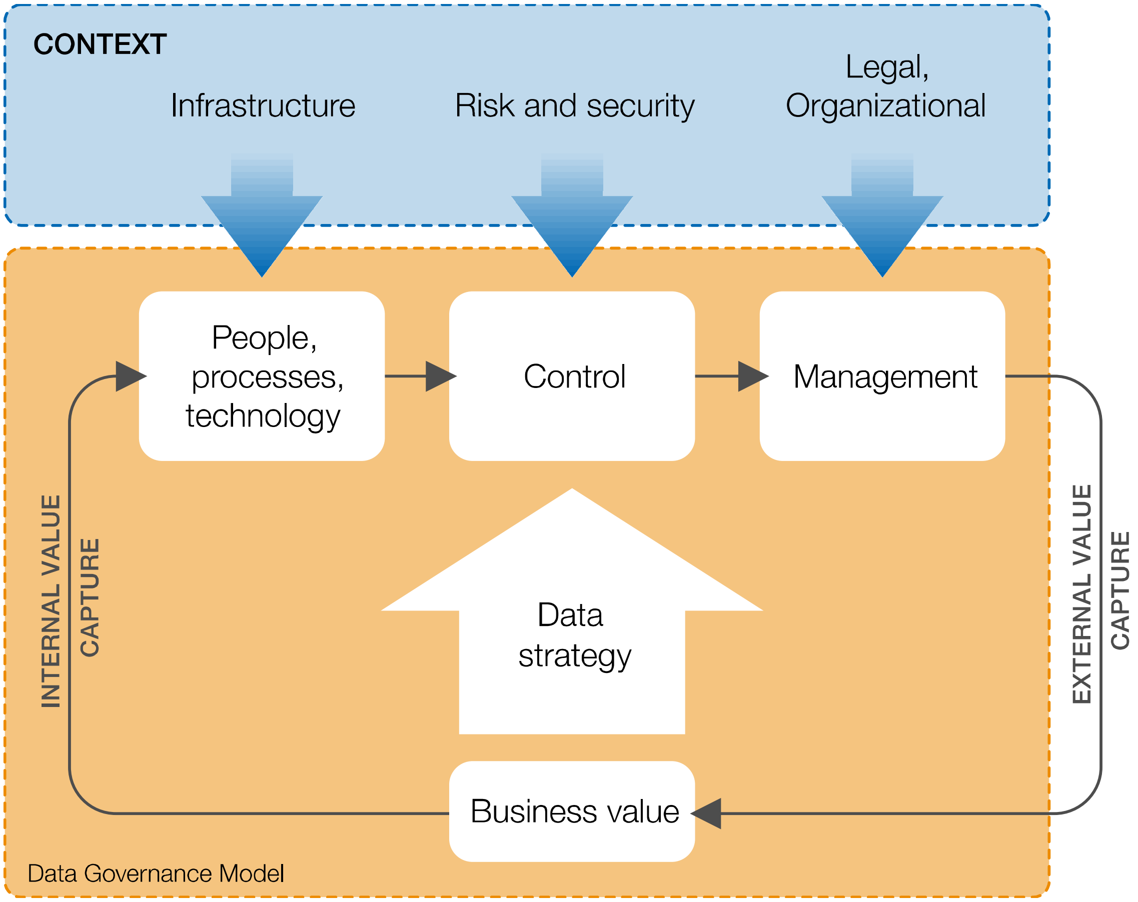


Figure 1. Data Governance model in context. Source: Authors

Figure *1* depicts the scope of the document in terms of internal and external value capture as well as the relationship of the model to its context. The key elements of context directly supporting the data governance model and discussed in this guidebook are:

- **infrastructure** that provides basic information technology tooling for manipulating, storing, and transferring data,
- **risk and security** that links the model with the safety requirements as well as dangers commonly linked to data processing and
- **legal and organisational** context, that provides the model with organisational support. These context drivers are only discussed in this guidebook to the extent they shape and influence the data governance model and its implementation.

1.4. How to read the guidebook

This guidebook is intended as a domain-specific data governance handbook applicable in the European Union context. It builds on top of resources such as Data Management Body of Knowledge (Dama International, 2009) and other sources, referencing and re-using them to minimise duplication.

The guidebook is structured based on the methodology described above and depicted on Figure *1*. In this approach, **data governance is seen as a feedback loop creating value for the organisation** through gaining control over the data an organisation handles and directing that value back at fulfilling a data strategy. Control over the data is enabled by people, processes, and technology and, in turn, enables data to be consciously managed. Data management leads to value creation enabling further investments into people, processes, and technology of data governance.

Explanation of the elements of the model form the first part of the guidebook with dedicated focus on data strategy and business value. Although the following sections can be read in isolation, familiarising oneself with the first part not only allows the reader to place the rest of the guidebook into a context but provides a mental model for thinking about data that can be used to solve problems not described in the current document. The first part of the guidebook seeks to answer the following questions:

- What function could data play in an organisation achieving their strategic goals, i.e., what is the business value data can provide?
- How to go about implementing that function i.e., what could a data strategy look like?
- What are the key interlinked concepts in deriving value from data i.e., what are the elements of implementing the data strategy?

The second part of the guidebook provides practical insight into how to implement the model in an organisation. In addition to explaining how a reader might utilise the model, the boundary between elements of the model and the surrounding context (in terms of structure and processes not covered by the guidebook) is given. The second part of the guidebook answers how an organisation would go about implementing described data governance model.

Both parts are divided into chapters. For an easy overview of the longer sub-chapters, we have created summaries and lead questions that your organisation could use for self-assessing the status of data management. The appendices contain tools for self-assessment and a model template for user´s consent on handling personal data.

2. Business value

Summary: *Business value can be derived from data and its reuse. Costs can arise from data obtainment, risk materialisation, risk mitigation, and resource management (infrastructure, people, improvement cycles).*

Lead questions: *How do you identify and generate value from data? How do you ensure that you will progressively eliminate inefficiencies and thus excessive costs? How do you ascertain a high level of data quality?*

In business, value is simply defined as the difference between benefits provided and costs incurred. The same definition is used by DMBOK to describe the value to be gained from data management (Dama International, 2009). This fundamental understanding should drive all data governance activities, regardless of the methodology used or the business context: data governance is not a goal to be achieved for the sake of it, rather should it drive tangible benefits that outweigh the costs incurred.

Value can be created in numerous ways not all of them being legal, desirable, or suitable in each context. It is the role of the data strategy, described in the next section, to define which benefits an organisation aims for and how it should go about achieving these. This in turn determines the types of costs involved.
In the following, let us look at various costs associated with data governance and the benefits that might be possible to reap.

The main types of costs associated with data governance are:

- The cost to **obtain** the data. Depending on the context, these costs could vary significantly. Collecting data from its users might be the main thing a start-up does but data can be by-product of some business process not necessarily geared towards data collection. The cost to design, build, market and operate an app collecting behavioural and physical data from the users is an example of the first type of cost and visitor data from a website is an example for the second.

  Bear in mind, that the costs of simply ending up with a dataset is but part of the total cost to obtain it. Typically, obtaining data also involves making sure the process is fully legal in the legal context of both the end users and the organisation obtaining the data and that the data is both semantically and technically suitable to be used for value creation.

- The costs of **risk materialisation.** When risks materialise, they, by definition, incur a cost for the organisation. However, most risk management frameworks do not deal with the costs but with the more general concept of risk impact. The reason for this distinction is, that risks can be existential in nature: when these materialise, the organisation ceases to exist and, depending on the legal structures used, stakeholders might be left legally or financially liable.

For example, leakage of sensitive patient data might lead to legal proceedings as well as revocation of the license to operate and lost datasets might not be possible to replace or recover. For these types of costs, the costs of mitigating the risk beyond reasonable likelihood should be used instead. When involving risk materialisation costs in the data governance cost structure, the cost of risk materialisation should be multiplied by the probability of risk materialisation within a given timeframe. For example, when something bad is likely to happen once every five years and cause 100 000.- worth of damage, its annual cost is $\frac{1}{5} \times 100000 = 20000$.

- The costs of **risk mitigation**. The second half of the risk-induced cost of data governance is the cost of mitigating the risks involved. With risk mitigation costs, two key things should be kept in mind. Firstly, the cost of mitigating the risk should not exceed the cost of risk materialisation multiplied by the chances of it happening. For the risk used an example above, for example, it would make no sense to spend more than 20 000.- in trying to mitigate the risk as it would be more efficient to absorb the risk materialisation cost Secondly, no risk can be entirely mitigated as mitigation measures create new risks or there is some residual risk elimination of which is prohibitively expensive. Therefore, the cost of data governance includes both risk mitigation and risk materialisation costs.

  Observe, that risks often involve secondary effects like impact on reputation, that might be difficult or impossible to convert to financial figures. This is especially the case in healthcare, where patient health and safety might be at direct risk. While direct costs of losing patient data might be relatively low, the impact it has on consumer confidence might significantly slow business growth and thus have a significant secondary impact.

- The cost of data **improvement**. While a dataset is stored in a computer system somewhere, the world around it keeps on moving and thus the dataset slowly but surely gets out of date. Also, investing into fixing data quality issues, enriching datasets with additional facts, or even gaining an understanding of the quality levels a dataset has might allow for more value to be gained from the data and thus might be worthy investments. All these activities incur a cost that, fortunately, are relatively easy to assess compared, for example, to the risk-related costs of data governance.

- Costs of **people, processes, and technology.** As implied by the data governance model described above, successful data management involves being able to control the data which in turn requires people, processes, and technology. What these might entail, is discussed in more detail in section 0 of the guidebook. Regardless, however, of what precisely is undertaken to gain control of the data an organisation processes, these activities require significant investment.

- **Infrastructure** costs. Obtaining, storing, processing, and destroying any electronic data involves computers doing work. These computers require capital investment to acquire, are picky about their physical environment and require human resources to manage properly. In the modern world, the infrastructure costs are either mostly or entirely hidden behind a bill from a cloud service provider. It can be lucrative, however, to look at the infrastructure used and ponder alternatives be it changing your system architecture, training operational staff, or switching providers which can all significantly reduce the infrastructure costs of data governance.

- **Data management** costs are potentially the highest cost factor of data governance and the ones most difficult to quantify reasonably. These costs are incurred through the process of actively managing the data to generate value. The main difficulty here lies in the fact that often data management is what an organisation *does* and so data management costs are deeply intertwined with the costs of running the business. It can be quite complex to determine, to what extent an activity is performed simply to keep the business running and to what extent should its costs be allocated to data management.

To cover the costs described, data governance should also deliver some benefits. These can be divided into direct benefits, where data is directly utilised for value delivery and indirect benefits, where data is used to aid another business process. Using heart rate variability data to calculate the recovery status of an athlete is an example of direct benefit data delivers while using the exercise data to determine if an athlete should be advertised biking or running shoes, is an example of indirect benefits.

Finding innovative ways to extract direct benefits from data is the crux of developing most digital business models whereas indirect value can be trickier to identify. The following list is by no means exhaustive but can serve as inspiration to find additional ways data can deliver benefits:

- **Waste** of limited resources or time is one of the key sources of inefficiencies of business processes. Often, these processes generate heaps of data that can be analysed to find and eliminate points of waste. Data can help improve the outcome by avoiding mistakes or serving as input for automation but can also underpin business process re-engineering efforts by providing insights into the process flow. For example, data on patient journeys through the healthcare provider can be used to identify unnecessary roundtrips and spots where better triage can help optimise the use of scarce resources.

- Identifying **stakeholder needs** is where data can provide the most secondary value as entire industries have sprung up around unmet needs identified using large-scale data analytics. Detailed analysis of customer behaviour correlated with other data sources can help uncover patterns indicative of unmet needs. For example, analysis of cases where people have not completed their treatment can indicate deficiencies in the treatment processes or uncover additional services that could help the patient stay on track.

- Better **decision-making** is an often-cited but complex area where use of data can be beneficial. While the relationship between raw data and human behaviour in general is not clear, the data-information-knowledge-wisdom pyramid (**Error! Reference source not found.**) is a common framework for relating data to higher-level decision-making. In this framework, data serves as a foundational element to a complex process leading to a wide range of inputs to a wide range of decisions from tactical to strategic and from deeply personal to impacting large organisations.

- All economic endeavour includes inherent **risks** and data can be hugely beneficial in both identifying risks as well as developing mitigating measures. In addition to economic risk, there is a safety risk involved in any non-trivial system from a piece of software in a smartwatch to an operating theatre bustling with human and technical activity. Here, too, data can serve to identify safety issues, predict their occurrence, and implement countermeasures. Finally, there is operational risk in any complex business process from fraud committed by partners to cybersecurity risks involving patient data. It is no surprise, that data generated by the business processes can be utilised to understand and mitigate these risks.

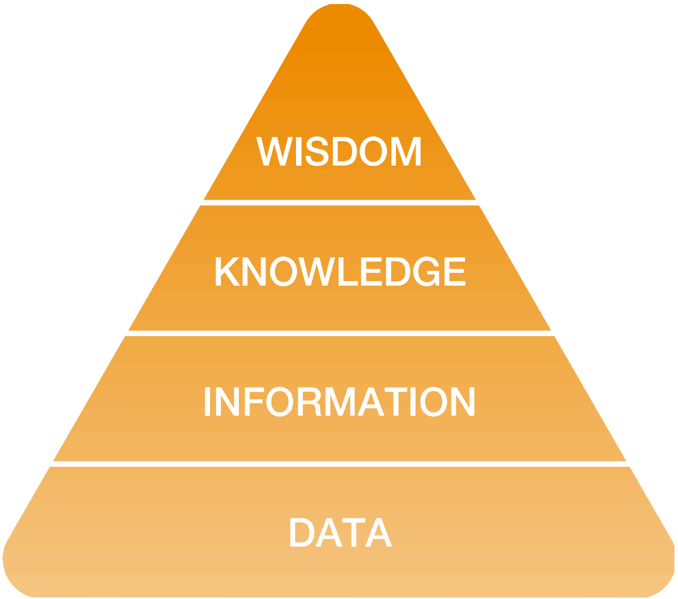


Figure 2. The data-information-knowledge–wisdom (dikw) hierarchy. Source: authors.

3. Data strategy

Summary: *Strategy is defined as a mental model that guides extraction of the given business value and is fundamentally seen as a set of brief principles guiding everyday work at the organisation. In the context of health services, the data strategy must be linked to the legal context of its implementation as well as the strategy and culture of the organisation implementing it.*

Lead questions: *What are the frameworks and regulations within which you will operate? How will you demonstrate compliance with them? How fast or slow is your ideal innovation cycle and how will your answer shape the structure of your organisation?*

An approach to strategy has been chosen for this guidebook that best fits the data governance framework used. In the model, business value determines the function of the activity, what are we trying to achieve. This is the role of the data strategy: to set forth the way the desired business value from data should be created. To illustrate this role, let us consider fever. There are many ways to deal with it from simply letting the body do its thing to cooling the body from the outside to using drugs to reduce the temperature. Which one is the most appropriate in a given situation is determined by the treatment strategy of the clinical expert.

Regardless of how an organisation thinks about data strategy, it needs fit into its context in two ways. Firstly, the strategy must fit the legal context - comprised of laws and regulations, contracts with stakeholders, constitutional documents of the organisation, etc. - an organisation operates in. For example, a strategy expressed as "obtain as much data as possible by concealing the purpose of its collection from the data subject" is very unlikely to succeed as the GDPR takes a rather dim view of such endeavours. Secondly, the data strategy must fit the organisation in terms of its culture. Peter Drucker has said that culture eats strategy for breakfast, meaning if a strategy runs afoul of the core beliefs and values of the organisation, it will not be executed upon regardless of the managerial pressure applied.

As with strategy itself, there are many approaches to articulating strategies. Data governance is a distributed and multidisciplinary activity by nature. Also, while data governance can be strategic in nature to the organisation, it relatively seldom receives attention from the top executives of the organisation. Thus, the person responsible for developing and executing the data strategy frequently lacks formal authority and relies on "soft power" instead. In this context, strategy should be possible to be distilled down to a few elements all stakeholders can agree upon and use in their work without much central coordination. Gadiesh and Gilbert describe such a model in their article "Transforming corner-office strategy into frontline action" (2001). In their view, it should be possible to distil a strategy down into a relatively small set of principles that:

- Forces trade-offs between competing resources. Example: principle "we do not own servers" forces a trade-off between the risks of cloud computing and the cost of developing infrastructure

- Tests the strategic soundness of a particular action. Example: principle "the legal basis of our data is rock solid" prevents an organisation from obtaining data from dubious sources, allows to validate usability choices when acquiring customer consent, dictates legal language in the terms of service etc.

- Sets clear boundaries within which employees operate and experiment. Example: the principle "we never risk the confidentiality of customer data" sets clear boundaries as to whether data can or cannot be anonymised, where it can be stored, how it can be handled and enables the employees to take responsibility for risks, for example, in terms of availability of customer data.

3.1. Basic principles

A data strategy should:

- Clearly state the time horizon an organisation defines as the "future". This provides the organisation with a clear sense of how far into the future they should be looking when planning infrastructure or making technical decisions. This is not to be confused with the time horizon of the strategy itself: a five-year strategy can define a one-year time horizon to emphasise flexibility as well as a ten-year time horizon to guide the organisation towards stability. Especially tools and technology choices can depend heavily on whether the organisation in question is a fast-iterating start-up, where everything will be rebuilt every six months or a mature business looking to capitalise on stable robust data infrastructure for years to come.

- Address the FAIR data principles in the context of the organisation clearly stating how each of the principles will be followed and to what extent. This relates the data strategy to the wider context of how health data is seen in a wider community of data consumers. FAIR data principles are a set of guiding principles to make data findable, accessible, interoperable and reusable (Wilkinson et al, 2016).

- Establish the risk appetite of the organisation in terms of data processed. This allows the internal stakeholders to have a well-defined shared notion of the level of necessary risk mitigation measures easing their implementation across the board.

- Define the regulatory context the organisation seeks to inhabit. While compliance to applicable regulation should never be an acceptable risk, organisations often can change their behaviour in effect changing, which parts of the complex global network of regulations they are subject to. For example, an organisation can decide not to collect data subject to especially draconic requirements, can limit their userbase, choose the physical location of datasets etc. This is especially true in the health and care domain, as health data is recognised as sensitive without a global consensus of what degree of sensitivity requires; while the GDPR has assigned it a “special category”, the legal status of health data and who can use it differs drastically around the world. Having clear boundaries would allow internal stakeholders to avoid accidentally overstepping boundaries yielding minor local gains at the expense of significant compliance impact.

- Define the desired process maturity of the data governance processes. High levels of process maturity (See **Error! Reference source not found.** below) can decrease flexibility and increase overhead while reducing risk and increasing predictability. Each organisation needs to clearly define their desired balance between up- and downsides of highly mature processes. For a fast-growing dynamic start-up low process maturity can be a valid strategic choice as innovation stemming from inherent unpredictability can be more valuable than risk reduction. For an established highly regulated provider of healthcare services handling large amounts of sensitive patient data on the other hand, a very high maturity level might be a prerequisite of obtaining the necessary certifications and licenses. The question of process maturity is closely linked to the time horizon in which the organisation operates: high-maturity processes seldom yield short-term benefits.

4. Data governance model

Summary: *Data governance is all about the process of managing various aspects of data within an organisation – from availability to value capture to security. In fields as complex as health services, the interdependent players need a strong data governance framework to help them direct their efforts in a purposeful manner.*

Lead questions: *What mechanisms can you develop to avoid under- or over-investing into areas like innovation and value capture? How does your data governance model have to evolve as the degree of data maturity inside your organisation rises? How big is your organisation's risk appetite and how does this affect your governance?*

The data governance model depicted on
Figure *1* consists of several interdependent elements creating a feedback loop that, when executed properly, should allow for a sustained growth in value derived from data management. **Internal value capture** directs business value generated by data management towards investing into **people, processes, and tools**. These in turn allow the organisation to take **control** of the data flowing it and, as a result, **manage** the data. Data management allows the organisation to create value for external stakeholders, that can be captured and directed towards further investment via value capture closing the loop.

It is important to note, that for the model to function, all its elements must be in place and the loop must close. Failure or success of each element has an impact on the usefulness of every other element. Data management without having control over it (including having knowledge of what data is there to be managed, being able to access or move it, etc.) is resource-intensive and can only yield limited benefits. Data control in turn requires the ability of an organisation to invest into the means of doing so.

Also, even if all the elements are in place and link together, the cycle can take time to gain traction. After all, in the beginning the organisation might not see much reason to invest into tooling or people and so the gains from early data management practices are low leading to only slow further investment into data management. Given the common annual budgeting cycles it can take several years before sufficient benefits from data management materialise and can be directed towards creating a serious attention to gaining control of data and extracting value from it. Thus, patience and readiness to wait out the slow first third of an exponential growth curve are necessary for the model to be thoroughly implemented.

As all processes require constant control, so does the process the model describes. Its behaviour and outcomes should be constantly assessed with adjustments being made as necessary.

4.1. Internal value capture

In management theory, two separate value-related processes are considered:

- value creation, that describes the way an organisation creates value by utilising the resources at its disposal

- value capture, that describes the process an organisation turns the created value into further resources

Value capture can in turn be separated into external and internal value capture with the former allowing an organisation to convert value created for outside stakeholders into input to the organisation and the latter directing the flow of that input within the organisation. Thus, internal value capture is a process by which business value is directed towards people, processes, and tools for data control. One can think of this as the process where some of the profits gained are invested into the manufacturing base in the hope of further increased profits.

The process is necessary, as typically value gained from data management materialises in a different part of the organisation than sees the budget line of expenses allowing for that value to appear. For example, the development team of a popular mental health support app might see all the revenue from app sales and be tempted to invest all of it into gaining additional app revenue while forgetting the data science team working on the AI powering the app.

Also, naturally, a business might both under- and over-invest into creating traction in the data management area: it might seek to invest into other business endeavours or tap additional sources of capital to boost strategic progress in the data management field. That said, a sustainable data management model keeps a healthy portion of the value gained from data constantly flowing back into maintaining and improving data management elements while also generating surplus value for the organisation in general.

4.2. Data management elements

Summary: *This section details the three key elements of data management: people, processes, and technology. It ties together topics as diverse as business culture and values on one hand, and legal processes and technical tools for data management on the other.*

Lead questions: *Are your organisation's and employees' culture and skillsets conforming with your internal processes? Which tools can help you create a shared understanding of the tasks at hand, e.g., a business vocabulary handbook? How do you plan on improving your data governance maturity over time? What kind of data tooling will your business model predominantly require and how do you intend to sustain it?*

4.2.1. People

The first and primary prerequisite of gaining control over the data of an organisation is having people who have the necessary soft skills – gained through personal development – and hard skills – acquired through study or training to do so and are also motivated and organised to perform the related tasks. A well-led group of motivated and smart individuals moving towards a shared goal can develop the necessary processes, adopt tools and technology, or even develop necessary skills. Adversely, no number of processes and tools can make a weak team perform truly well.

There are four key areas to consider when developing people in the context of data governance:

- **Culture and values**. To be able to devise and execute a data strategy in terms of people, processes and technology, a set of shared values and a strong positive culture is very much necessary. Such a culture would often reward employees and create an environment where they can develop and operate at their full potential. There is, however, no single kind of culture that would guarantee success. Every organisation has their own strategy, their own history, and their own goals: good culture is whichever culture helps the organisation in executing that strategy, achieving these goals, and coming to terms with the history.

- **Competences.** Data control and management assumes a level of both soft and hard competences that need to be actively developed. In general, an organisation engaging in a as complex of an endeavour as data management, should have a robust competence model in place describing functional and core competences needed for the execution of organisational strategy. To this model, the needs of data management add mainly functional competences of working with data and doing so safely and securely. Also, as dealing with data is commonly a cross-disciplinary activity, the ability to work well across disciplinary and organisational boundaries becomes important.

- **Organisation**. Competent people with aligned values need to be organised in a fashion that allows these competences to be meaningfully applied to gain control of the data and to manage it effectively. All these models can be useful if the choice is made and communicated deliberately while taking into consideration the overall structure of the organisation in question and the main business processes being executed.

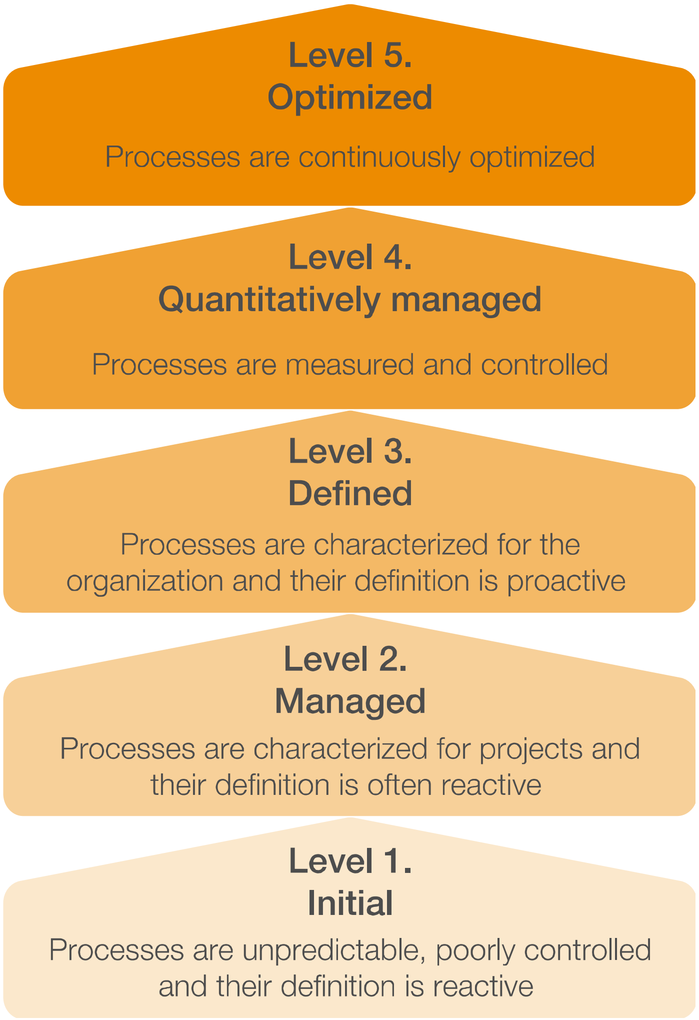
- **Leadership and management.** None of the elements described above are guaranteed to arise spontaneously when a group of people is told to "get a grip on the data". Data management in any organisation requires specific planning, leadership, and coordination in the sense of managing the people involved. In the ideal case, these form an integral part of organisational structure. In particular, the role of the leader is to assure strong positivity of the culture while assuring the team members involved have a realistic understanding of their competences with enough motivation and resources to develop these in the desired direction.

4.2.2. Processes

To organise the people involved in data governance in general and gaining control of data in particular, a set of processes is required. Typically, all organisations have the requisite processes in place to some extent, but these might not be suitably mature. In the software world, a capability maturity model is typically used to assess the maturity level of a given process: a higher maturity level indicates a more robust and adaptable process more likely to fulfil its role in the data governance model described in this document.

Higher process maturity is also associated with lower risks. (For example, there is a smaller chance of someone forgetting a laptop with patient data on public transport, when the process of handling said data is well-defined and strictly enforced.)

On the other hand, higher process maturity increases overhead and can reduce flexibility. It is therefore of paramount importance, that organisations have a clear understanding of both the current and desired maturity level of their data governance processes.



Level 5.
Optimized
Processes are continuously optimized

Level 4.
Quantitatively managed
Processes are measured and controlled

Level 3.
Defined
Processes are characterized for the organization and their definition is proactive

Level 2.
Managed
Processes are characterized for projects and their definition is often reactive

Level 1.
Initial
Processes are unpredictable, poorly controlled and their definition is reactive

Figure 3. Outline of the capability maturity model.
Originally developed at Carnegie Mellon University and now administered by the CMMI Institute

In general, data governance processes can be divided into three main groups:

- **Organisational processes** focused on the way the organisation interacts with the data
- **Data-focused processes** that deal with the data directly
- **Metadata process** that handles data about data

Organisational processes enable data-focused processes by providing a conducive context for them to perform. Also, the metadata process applies both process groups to the data generated by data-focused processes: metadata is also data requiring a set of data-focused processes supported by organisational processes and generating its own metadata.

The main organisational processes are:

- **Legal** processes, that assure the legal basis for holding, sharing, and processing data is legally sound.
- **Risk management** processes, that assure the risks stemming from data governance are suitably managed and monitored with mitigation measures and incident response plans frequently tested.
- **Compliance** processes, that assure the data governance implemented by the organisation is compliant with the relevant regulation.
- **Data control processes,** that revolve around actively managing elements of data control (See 4.4).

These organisational processes focus on making sure things happen rather than making things happen. Risk management is about knowing and monitoring the risks involved, not mitigating them, for example.

Data-focused processes depend on the lifecycle stage the data is in. The data lifecycle provided by DMBOK is depicted on Figure 4.

The main data-focused processes are:

- Data architecture describes the structure of an organisation's logical and physical data assets and data management resources. **Data architecture management**, therefore, encompasses the process of actively making decisions on the elements of this structure as well as their relationships.
- **Data quality management** provides a context-specific process for improving the fitness of data that’s used for analysis and decision making. This process generates metadata on the quality of data items and uses these as input to various activities to improve data quality.

- **Data tracing** is a process for generating metadata on data lineage (i.e., the sources and sinks of data connected by paths data items take through the organisation) and lifecycle (i.e., what are the stages data items go through between being acquired or created and being destroyed).

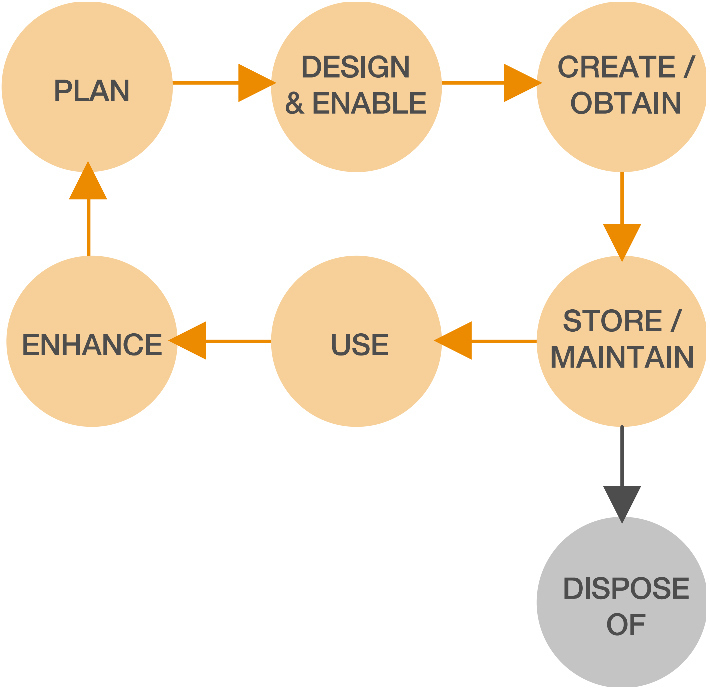


Figure 4. Data lifecycle key activities. Source: Dama International, 2009

4.2.3. Tools and technology

Firstly, to implement any tool correctly, the business process vision behind it must be understood. Although tools can have many uses, chances are a tool is best used for what it was designed for. In the fast-developing world of data management, many available solutions are built by an organisation to solve their specific problem in a specific fashion, so the tools become opinionated.

Secondly, avoid the downsides of combining adoption of standard tooling and development of custom solutions - adopting a best practice and then forcing it into a shape the organisation decides it needs. This negates all the effects of the best practice embedded in the tool while incurring costs of custom solutions without the flexibility.

There are four main drivers that define the nature of the tooling an organisation might need for data governance:

- **The level of control desired**. As explained below, there is no single level of "control" over data and the decision "to what extent do we need to be in control of our data?"

is of strategic nature. However, that decision has a direct impact on the tooling implemented. An organisation with minimal control needs will also need minimal data governance tooling while an organisation seeking near-complete control over their data assets is going to require a lot more sophisticated toolchain.

- **The processes in place and their maturity.** To achieve highly mature processes, these processes need to be measured and adjusted. This creates data and is difficult to achieve without dedicated tooling. Also, data-focused processes typically require their own dedicated tools to be executed**.**

- **Data management needs.** Direct value creation often also has specific needs towards data control tooling. For example, data analysis in conjunction with strict privacy rules creates a requirement for a complex cryptographic privacy-preserving data analysis framework while the need to manage a large set of consents from data subjects requires high-quality data lineage tooling linked to consent management.

- **Data strategy.** Different aspects of the data strategy from risk appetite to time horizon directly influence the choice of tools and technology for data governance. Time is a particularly important factor as the trade-off between being able to move fast and being able to rely on a solid foundation is particularly difficult in the context of tools and technology.

Tools and technology for data governance can roughly be split into the following three categories:

- **Process tooling**, that supports data governance processes. These tools support data architecture management, data quality management, data mapping, metadata management etc. processes and can range from simple spreadsheets to sophisticated software solutions tailored to the organisation. Collaboration platforms like Confluence or Sharepoint are a good example of process tools suitable for many kinds of processes from knowledge management to defect tracking while a dedicated tool like SAS Metadata Server is focused solely on metadata management.

- **Data access tooling**, that support storage and processing of data. These tools solve the technical obstacles of storing large amounts of data in line with the risks present but also allow this data to be accessed in ways the organisation needs. Databases, be it relational, object-based, or graph-driven are good examples of data access tooling.

- **Data protection tooling**, that supports protection of data processed by the organisation. These tools typically are integrated to either process tooling (data access rights, for example, are part of metadata) or data access tooling (generation and lifecycle of API access keys for external data access or handling encryption of data at rest, for example). There are, however, cases, where data protection requires dedicated effort that is separate from both process support and data access. For example, encryption solutions perform sophisticated manipulation of data that serves the sole purpose of data protection rather than supporting a particular business process.

4.3. Control

Summary: *Management of any object requires the ability to control that object, it can be quite hard to sheer a sheep without being able to catch or hold it in place. This is especially true for something as impalpable as data. This control is supplied by the people, processes, and technology an organisation possesses.*

Lead questions: *How is data storage and manipulation handled and who is responsible? Which data boundaries are necessary or helpful to establish both within the organisation and in collaboration with external partners? Are different roles and responsibilities related to data control over its lifecycle defined?*

Having control of the data means the organisation has the following capabilities regarding data:

- **Create, read, update, and delete.** This provides the organisation with basic tools required for any data management activity.

- **Move** data between places of storage. This allows the organisation to pick the most suitable storage location of the data both in terms of technical requirements like availability and accessibility but also from the perspective of legal jurisdiction (See also 5.5.3 of the Guidebook).

- **Convert and transform.** This allows the organisation to shape the data into a format most suitable for value extraction by, for example, either removing or adding personal identifiers or enriching datasets by linking them together.

- **Decide on data access.** This allows the organisation to fulfil their compliance requirements (in case of regulated data) and provides the basic tools for monetising data access.

- **Control over business processes creating and modifying the data.** Data is always generated by a specific business process, sometimes as a by-product. This data creation process determines the initial quality and properties of the data, and it is difficult to achieve full control over data without the ability to influence the business processes it results from.

There is no such thing as an absolute control over data. After all, there will always be legal or practical restrictions on moving data about and there is an ever-present risk of a security breach denying the organisation the ability to control access to the data. Also achieving and maintaining control over data requires resources.

It is therefore important to understand the level of control required to perform the data management necessary for value delivery and set it forth as a set of realistic meaningful goals. For example, a very strong level of control over data processed might be a prerequisite for even gaining access to medical data while the healthcare processes and devices generating the data might be entirely out of control of the organisation in question due to them being operated by the healthcare provider.

The concept of control over data is intimately tied to the concept of risks. After all, information security risks mainly deal with the elements of data control: being able to access data (availability), being able to control data access (confidentiality) and the ability to control the way data is modified or created (integrity). Therefore, a mature risk management capability is a prerequisite for adequate assessment of the level of control an organisation has over their data.

In general, control over data assumes a level of control over the data-related processes of the organisation: the higher the maturity level the higher the level of control.

Achieving control over data assumes deep knowledge about the data. This includes but is not limited to what is traditionally considered metadata. In particular, the following additional knowledge of data is necessary:

- **What data is processed or stored where and how**. This implies control of both functional and technical architecture of the organisation and a deep knowledge of the business processes generating the data.

- **Organisational and data boundaries.** All organisations interact with their surroundings and these interactions commonly involve exchanging data with the outside world, be it the invoices or reports sent to customers or complex machine-to-machine interactions with national healthcare systems. It is recommended to have an accurate boundary document describing all the ins and outs of the organisation from the perspective of data.

- **Data lineage and lifecycle.** As data enters the organisation, it is continuously transformed, moved around, and combined with other data elements creating a complex network of data lineage from which finally value emerges. Knowledge of that lineage network is a prerequisite of being able to control it. In addition, data elements go through a lifecycle that, while containing main elements of acquisition, improvement, and destruction is unique to each organisation in terms of details. Control over data is hard to imagine without control over these lifecycle stages. Data lineage and lifecycle often intertwine to the extent that it is difficult to discuss one without another: data elements might cease to exist or be useful after being combined with others, improvement of datasets often involves changing its lineage, etc.

- **Legal context.** Data as an abstract concept is not commonly subject to legal restrictions. Personal and medical data, however, are commonly subject to severe legal requirements in terms of its processing or even location of the physical devices containing the data. Also, in a complex data ecosystem data often either enters or leaves the organisation under specific legal conditions stipulating requirements to its processing, storage, access, etc. This means having knowledge of the legal context surrounding each data element processed and the ability to follow that context through data lineage are crucial elements of data control.

- **Quality.** What constitutes data quality depends on uses and nature of the data. For example, blood type of a patient needs to be very accurately recorded and transmitted for healthcare to be safe whereas the address of the patient has little quality requirements for marketing purposes. The same blood type data, however, has much lower accuracy and quality requirements for general data analysis and the address could be required to be very accurate for use in emergency dispatch context.

  Quality can mean trustworthiness, completeness, accuracy (i.e., reflecting reality), adherence to an assumed statistical distribution, knowledge, and level of noise present etc. It is therefore crucial, that the organisation has a clear understanding as well as control of both the actual and required quality levels of the data it holds. This understanding of control is tightly related to knowledge of data lineage, as often data quality is determined at the source.

4.4. Management

Summary: *Data can be managed to create value in a variety of ways and there are different methods to do so. The relationship between data management and risk management is introduced as value creation which is inherently laden with risk. The concept of data quality management is introduced as a key element of data management.*

Lead questions: *How do you intend to classify the different kinds of data that you will process? What is required to make sure that all datasets are treated in accordance with their classification?*

The DMBOK provides the following definition of data management:

> *Data management is the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles (Dama International, 2009).*

This definition fits well with the approach taken in this guidebook as all the previously described steps of the model seek to provide control enabling data management. The people, processes and tools required to establish control over data create an understanding of the object to be managed as well as establishing the capabilities to do so.

Most acts of data management involve moving data through its lifecycle stages. Different types of data have different lifecycle characteristics and therefore require distinct approaches to their management. These differences can stem from both legislation as well as nature of data. Article 9 of the GDPR, for example, lists 8 different kinds of data (including health data) and describes limitations to their processing. National regulations can have different definitions of data types.

As data can transform from one type to another during its lifecycle (e.g., personalised medical data gets anonymised to be analysed with results published as open data), correct management approaches are applied as these changes occur.

Some main data types relevant to the field of active and healthy ageing are:

- **Health or genetic data** is personal, highly sensitive, strongly regulated in both the GDPR as well as national acts and has high requirements in terms of accuracy. Such data typically enters and leaves the organisation in a very controlled manner being both acquired and handed over or destroyed under legally well-defined circumstances. Management of medical data therefore requires high degree of control over it which, in turn, assumes high maturity levels from both the processes in place to make the data possible to manage as well as the management processes themselves.

- **Personal data** is also personal and highly sensitive but is typically less strictly (and differently) regulated, than health or genetic data. In general, data of this type also requires attention to detail in terms of both acquisition and deletion, as health or genetic data and has similar requirements towards process maturity. However, as national regulations differ in terms of data classification, it is important to ensure the dataset in question is not subject to special regulatory concern.

- **Derived data** is data, that is created by processing other types of data. Derived data includes but is not limited to data created by anonymisation, pseudonymisation or de-anonymisation, results of data analysis, metadata, datasets created by combining multiple datasets etc. The key challenge with derived data is to ensure it is correctly categorised and treated as such.

> *For example, analysis of heart rate data (health data) would allow one to derive the number of days a person engages in sports (personal data) whereas the latter can be further be processed to yield mean number of days a person in an age bracket in a country engages in sports per week (could potentially be treated as open data). Also, as derived data is created by the organisation and not obtained either from a data subject or a third party, its processing requires a legal basis either defined in the consent given or the legal framework surrounding the data transfer. The purpose of data processing, as listed in the consent given by the data subject, must cover creation of the derived dataset.*

- Data is **commercially sensitive** if loss of any of its security attributes (See **Error! Reference source not found.**) will lead to financial loss for the organisation. The key difference between commercially sensitive data and other data types (although personal or medical data can certainly be commercially sensitive) is the process focus.

Here, the focus is on risk management in general whereas with other types of data the focus is on managing compliance, oftentimes seen as a sub-activity of risk management.

- **Open data** has many overlapping definitions but in the current context it means data that is publicly available. As opposed to other data types, where eventual destruction of data at the end of its lifecycle must be ensured, open data should commonly be prevented from being destroyed. Also, open data management involves additional aspects like ensuring the license terms are both appropriate and up to date, assuring the technical availability of the data, updating the documentation etc.

Multiple frameworks exist describing the various activities involved in data management. These tend to focus on the level of detail not specific to any domain and are thus outside of the scope of this guidebook. The DAMA-DMBOK2 Data Management Framework, depicted on Figure 5, is one of such frameworks. It splits the field of data management into 10 separate knowledge areas, each connected to the central concept of data governance.

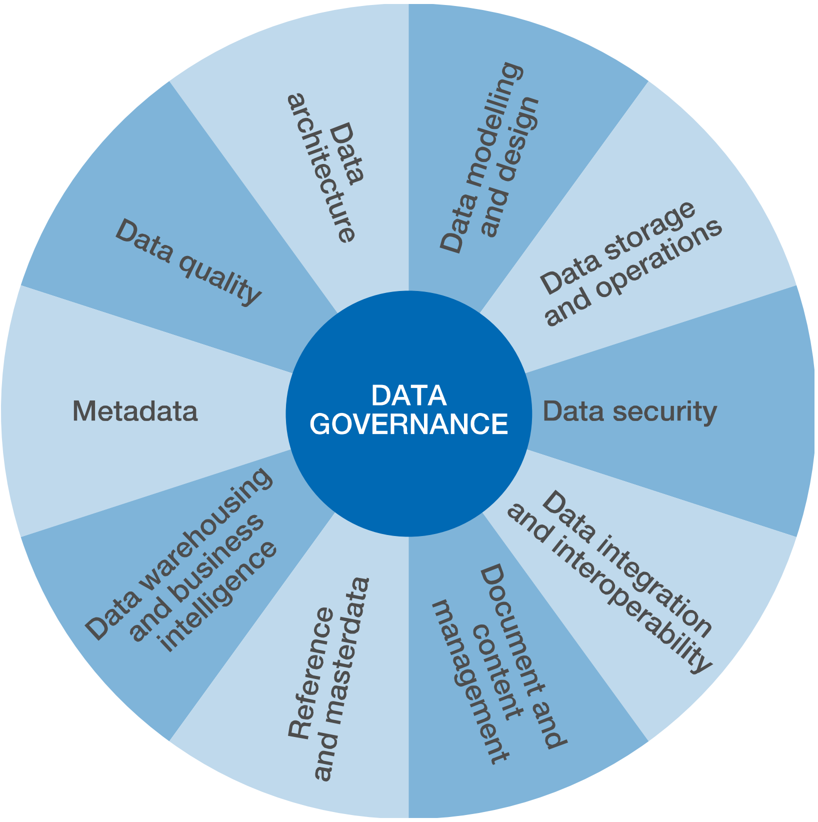

DATA GOVERNANCE
Data architecture
Data modelling and design
Data storage and operations
Data security
Data integration and interoperability
Document and content management
Reference and masterdata
Data warehousing and business intelligence
Metadata
Data quality

Figure 5. Data management framework. Source: Dama International, 2009.

Regardless of the data management framework used, two key areas stand out as relevant in the active and healthy ageing area: risk management and quality management.

From one hand, health data is highly regulated (see Chapter 5.5.4) and thus highly risky from compliance standpoint. On the other, any value creation is inherently linked to risks. Also, both potential value of data and the risks related to it are highly dependent on the quality attributes of data. Indeed, data quality is defined as the extent to which data is fit for a purpose while risks are oftentimes associated with loss of certain quality attributes of data. Having higher-quality data makes the dataset both more useful and more risky, as higher-quality data is more useful for both legitimate and illegitimate purposes. Lowering the data quality, however, will both reduce its usefulness and create risks e.g., through incorrect treatment of patients.

Thus, the balance between managing risk (including compliance risk), managing quality and creating value is central to the very business model of an organisation in the active and healthy ageing field. Achieving that balance requires in-depth knowledge of both relevant areas of data management as well as business practices and overall strategy of the organisation.

5. Implementing data management

This second part of the guidebook explains how to practically implement the model described previously explained. All sections in the chapter contain lead questions to enable readers scan the current state of their organisation in a given area.

5.1. Internal value capture

Summary: *The section provides guidance on how to budget for technology and people for data governance. Relationships between CAPEX and OPEX are discussed in terms of maintaining the systems and the staff involved in operations.*

Lead questions: *Which kind of CAPEX/OPEX relationship does your business model demand? What is the risk appetite and resilience of your organisation, and does it match your investment strategy?*

The key goal of internal value capture is to direct a portion of the value captured externally towards developing further capabilities to create that value. The costs and benefits of creating value from data do not necessarily appear together and, therefore, a conscious effort is needed to direct funds towards developing data management elements. The internal value capture process is mostly about corporate finance and is thus comprehensively covered elsewhere. There are, however, two ideas specific to data management that could help make sense of this process and avoid pitfalls.

Firstly, let us consider the relationship between capital expenditure (CAPEX) and operating expenses (OPEX). Capital expenditure denotes one-off investment into things an organisation intends to utilise for long periods of time while operating expenses denote running costs of an organisation like salaries, office costs etc. In case of material assets, the assets are bought using CAPEX, incur only financial amortisation during their use and are finally replaced by new CAPEX. Whether the investments occur every time equipment is replaced or are spread more evenly over time is a question of finance management. In case of immaterial assets like software and data, however, this relationship is much trickier.

As described previously, data needs to be actively managed to yield value and the same is true for software. Thus, a capital investment into implementing new software for data management, acquiring a new data source or, indeed, creating new software creating new streams of data has a direct impact on the operating expenses of the following periods. Software needs to be updated, it requires monitoring it breaks and needs to be fixed, the needs of the customers change constantly over time.

The same is true for data: the processes described above (See 0 and 0) require people to execute and might incur other costs as well. It is therefore not advisable to invest heavily into data management elements without a clear source for the operating expenses. It is very common for organisations to underestimate the costs involved leading to poorly managed datasets and software tooling that either do not yield a positive revenue stream or even act as a cost centre.

For software, a good rule of thumb is to spend annually about 20% of the original investment. This allows the entire system to be re-built within five years, this roughly aligns with a typical useful lifetime of software. Another approach is to forego CAPEX altogether and establish a stable support team incurring OPEX right away. This avoids the potentially dangerous transition from building to maintaining and assures the system is well-supported from the get-go.

For example, instead of investing 500.000 into building a new system, a development team with annual OPEX of 100.000 could be established right away and expected to release the software as it matures. As data is much more divergent in nature than common software systems, no such rule of thumb can be given for data. Thus, caution is advised when taking on new datasets unless solid prior knowledge of costs of maintaining such datasets in the current setting exists.

Secondly, investment strategies into human capital, the second large pillar of the data management elements, should be considered. One possible set of strategies to choose between is presented below (Philips, 2005). From amongst these, one should be picked and aligned with the overall human capital investment strategy of the organisation:

- **Let others do it.** Bring on board fully trained experienced team members not investing in their development. Replace people as new competences are needed or people seek to develop themselves. **For a knowledge- and data-intensive organisation, this strategy is unlikely to be sustainable.**

- **Invest the minimum**. Bring on board fully trained experienced team members investing bare minimum in their development to assure required staff turnover figures. **This strategy is more sustainable, than reliance on investments by others, but assumes availability of required competences at reasonable cost.**

- **Invest with the rest.** Monitor the market situation and assure to invest into human capital on par with the market. **This helps the organisation align itself with competitors while also creating little incentive for people to leave due to lack of investment.**

- **Invest until it hurts.** Go all-out in investing into human capital hoping for a pay-off in terms of externally captured value. **This strategy assures the organisation has a clear edge in the market in terms of human capital but assumes a relative lucrative business model or continued investments.**

- **Invest as long as there is a payoff.** Monitor the situation and invest carefully into select areas of human capital in line with revenues created. **This strategy assures the organisation is not over-investing maximising return on investment but assumes a potentially unrealistically accurate view of both immediate and strategic potential of human capital.**

5.2. Data management elements

Summary: *This section seeks to give guidance in implementing specific technologies, processes, or organisational capabilities to allow for gaining control over data. The processes and technology categories and main types listed previously are discussed in some detail along with practical implementation guidance. Also, practical advice is given on how to structure the technology investment in a way that aligns with the architecture of the organisation itself.*

Lead questions: *How are different roles and responsibilities related to data processing over its lifecycle defined? How do you ensure the integrity of data management processes? Which data quality metrics are of particular relevance to you and how do you improve them? What role does metadata play in your organisation and how do you formalise its management?*

In terms of organisational capabilities, **an internal stakeholder** model is presented with different roles like data steward, data protection specialist etc. listed, linked to data lifecycle stages, and supplied with required competences.

5.2.1. People

The leadership aspect of establishing people capabilities to achieve control over data is out of the scope of this document as it does not differ fundamentally from the task of establishing competences and teams for any other purpose. Leadership is also, to a very great extent, a matter of personal style and thus difficult to give specific advice on. That said, there are a few particulars about managing people in the data management context that need dedicated attention.

Firstly, the model presented on
Figure *1* is inherently feedback-based: value created is captured and directed towards creation of more value. Also, health care in general and healthy ageing in particular is a dynamic area where new methods and ideas immediately create a need for next ones. Organisations effect a change on their environment and the environment changes rapidly in response. This new environment needs a different strategy, different processes, and different competences to strive in.

Thus, conversely, the better people are at something the faster they need to be good at something else. The organisation's ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments is called dynamic capability and any organisation dealing with data in the active and healthy ageing area would do well to deliberately develop these capabilities (Teece at al., 1997).

Secondly, as data moves through its lifecycle, it requires different competences. Sharing data assumes different competences (developing an API, working out license terms, monitoring usage) from acquiring it from customers (developing a user interface, working out consent management processes) from destroying data (deep knowledge of hardware or cloud environments). Furthermore, importance of various roles (both legally mandated and required by the business processes) changes along with the changing needs for competences. To maintain this complex setup, an organisation should seek to maintain a good understanding of which people with which competences are fulfilling which roles even when the organisational culture leans towards less formal and less centralised person culture. As the number of both roles involved and competences needed with the growing business complexity, so does the number of people who focus on data management.

Various sources list roles related to data governance and management, both technical and non-technical. As a rule of thumb, all roles should provide more value than is the cost of manning them plus the cost of complexity they create in the organisation. While complexity tends to grow exponentially alongside team size, complexity-related costs tend to grow even faster. Ultimately, however, there is no objectively useable method for measuring and anticipating this area of organisational costs. Special caution should be exercised around "architect" or "administrator" roles as such roles tend to effect change through other roles which makes assessing their applicability difficult.

Combining regulatory requirements with best practice and practical needs of data governance, the following list of roles is likely to cover the necessary functions. Roles do not have to align with people: a person can successfully execute multiple roles (within reason) and one role can be implemented by several people.

- **Data steward** knows everything about data and is responsible for it. Typically, the data steward is also a subject matter expert representing an understanding of value created.

- **Data srchitect** designs systems, software and processes for data processing and relates the data architecture to the overall enterprise architecture.

- **Compliance officer** (also Data Protection Officer) is responsible for ensuring all data processing in the organisation happens in compliance with the relevant regulation.

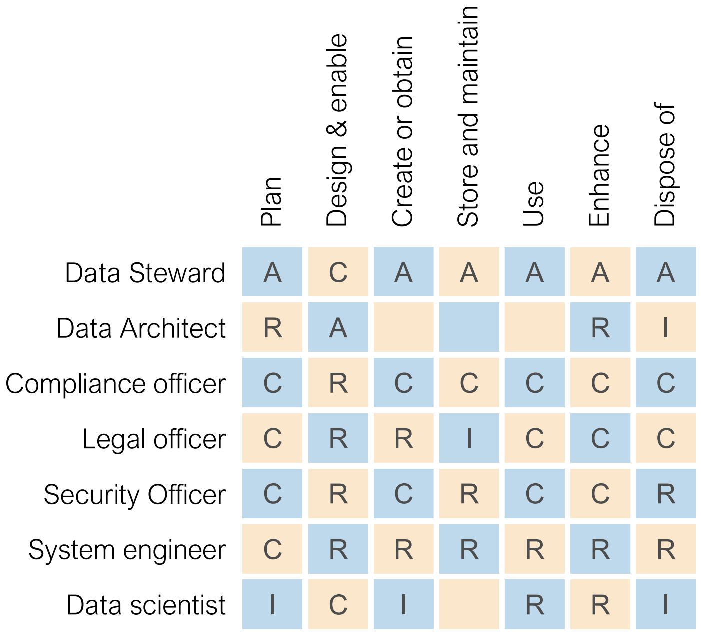
- **Legal officer** is responsible for ensuring all data-related activities have a solid legal foundation.

- **Security officer** is responsible for ensuring availability, confidentiality and integrity of the data based on agreed-upon levels.

- **System engineer** is responsible for building and operating software and hardware systems processing the data.

- **Data scientist** is responsible for the mathematical aspects of data processing.

Error! Reference source not found. relates the data-related roles to data lifecycle stages (See Figure 4) using the RACI model, where ***R*** denotes responsibility for data in this stage, ***A*** denotes accountability for making sure all data-related activities related to the stage are properly executed, ***C*** denotes a consultancy role and ***I*** a role that is informed about activities related to data at a given stage. The table is driven by formal definition of the roles and, of course, practical responsibilities of the roles can differ. It is important, however, to ensure all stages have exactly one role accountable for it.



	Plan	Design & enable	Create or obtain	Store and maintain	Use	Enhance	Dispose of
Data Steward	A	C	A	A	A	A	A
Data Architect	R	A				R	I
Compliance officer	C	R	C	C	C	C	C
Legal officer	C	R	R	I	C	C	C
Security Officer	C	R	C	R	C	C	R
System engineer	C	R	R	R	R	R	R
Data scientist	I	C	I		R	R	I

Table 1. Responsibilities of data-related roles towards data lifecycle stages. Source: Authors.

5.2.2. Processes

5.2.2.1. Organisational processes

For legal processes the key recommendation is to start establishing these as early as possible in the process of setting up the data governance framework described in this guidebook. Quite often, legal processes are seen as a necessary evil to be bolted on to existing business processes. This will lead to frustration on both sides as well as diminish opportunities in process development. As described above, a successful data governance setup is an optimal balance between multiple conflicting factors. Since many of these factors are of legal nature, designing legal processes alongside the other data governance and management processes with both supporting each other would benefit finding that optimum.

Oftentimes, all that data-driven organisations do is data governance. However, the data governance model this guide describes can also be operated by a smaller section of a large organisation: a hospital might have a department offering data-driven care services or an insurance provider might have spun off their data analysis department. In such cases, it is highly recommended to link the risk management processes of the smaller organisation with that of the larger. A small team might not have the resources or the competences to engage in thorough risk management and the larger team can be tapped for support. And the risk folks of the larger organisation are likely to be happy to have a part of their risk portfolio under better control.

Regardless of how and by whom risk processes are designed, they should always involve regular testing of risk mitigation measures. Data flows of a dynamic organisation can easily be, at least partially, out of control and predicting which parts of the complex system are critical to the functioning of the whole is also tricky. Therefore, a routine testing procedure should be in place to test the system in various failure modes to understand, what parts of the data flows are not well maintained and whether these are critical to the overall system behaviour. Two types of testing are highly recommended:

- **Penetration testing**, where an external actor attempts to gain access to key systems of the organisation in live production environment. This testing method is preferred over static code analysis, system analysis and other methods (that of course have a purpose and should be performed, as necessary) as it provides concrete evidence of significance of vulnerabilities rather than a theoretical argument that can easily be contested.

- **Recovery testing,** where loss of either the entire information system or key parts of it is simulated with recovery attempted according to the pre-defined plan. Mere existence of a backup is not sufficient if the backup is stale, cannot be accessed, does not include required parts of the system, or takes unreasonable effort to deploy. Again, the goal of this testing method is to replace theoretical musings of a plan with concrete evidence of the crown jewels of an organisation being recoverable in case of catastrophic failure.

The risk processes in place in a data governance context should consider the inevitably high complexity of the system at hand. For complex systems human error and unfortunate alignment of several minor failures is a common source of a catastrophic failure. Also, such systems drift into unsafe states as the risk control processes themselves decay or separate from the reality over time. To handle such systems, a safety model called STAMP (Leveson, 2016) has been developed and should be taken into consideration when designing risk processes for data-intensive complex systems.

Compliance processes stem from a strategic decision an organisation makes about what they are compliant to. As stated before, being non-compliant to a regulation is not usually a worthwhile risk to be taken. But the organisation certainly can decide if they want to engage in activities subject to regulation. Among the vast array of decisions that will inevitably impact the compliance exposure of an organisation are questions such as which infrastructure provider to use, which server locations to use, which customers to serve, where to incorporate, what data fields to collect etc. Whereas in many organisations these decisions grow organically out of product decisions, they should be made explicit and clearly acknowledged as having a strong strategic impact.

An organisation can also stipulate compliance-driven limitations to activities in their business or data strategy. Another approach is to limit expensive and complex compliance exposure to a part of an organisation. One can for example imagine a dedicated subsidiary of an organisation receiving and analysing patient data and presenting the rest of the organisation with processed results that are sufficient for the core business of the organisation but are not subject to audits, heavy supervision, certification etc.

In an international environment, the question of regulatory priorities often arises. Even in EU, where data protection is regulated in an overarching manner, protection of health data is regulated by each country individually. It cannot therefore be guaranteed that any combination of the data subject, physical processing location, incorporation location, healthcare provider legal status and location etc. can be clearly resolved to a single regulation of a country. Thus, it is possible for different regulations to be in conflict. And, therefore, a clear process as well as capabilities must exist to make an informed strategy-aligned decision about which regulation to prioritise.

Even if various regulations are not in conflict, it can be complex to track, which data elements in which stages of their lifecycle are subject to which regulations of which countries. Therefore, it is advisable to link compliance processes with data mapping allowing to assure compliance based on not only data type but also attributes of data lineage such as location of data subjects, source of data, physical location of servers involved, types of operations performed etc.

Data control processes are basically about making decisions based on metadata and about metadata effectively answering the question of whether metadata indicates data to behave in an optimal manner and do we know enough about our data to make that decision. One approach to data control is to focus on system boundary having processes in place to maintain a clear understanding of all the ways data enters and exists the organisation. This allows a good perspective on data within the organisational perimeter in turn allowing to establish processes to control it.

In achieving control over data, risk management can be a useful and well-described resource if clear understanding of the level of control over data exists and is provided as input to risk process design.

5.2.2.2. Data-focused processes

The two main data-focused processes are **data architecture management** and **data quality management**, both of which will be discussed below.

Data architecture defines the blueprint for managing data assets by aligning with organisational strategy to establish strategic data requirements and designs to meet those requirements (Dama International, 2009). This implies a separation between the model (i.e., the description or blueprint) and the actual implementation (i.e., the way data assets are managed). However, making sure that the blueprint does not differ too wildly from what happens in reality is one of the most important tasks of any architect.

Whichever definition is used, the approach taken should not differ significantly from the way the architecture of the organisation is thought about by decision-makers. Although communication is important, the main reason for this is, that it is hard to make conceptually different model align but align they must. Conway's law states that organisations design systems to reflect their communication structure (Conway, 1968). Therefore, designing data architectures not in line with how the rest of the organisation operates will create tension to be released either by implementing a different data flow (potentially causing a departure from the blueprint) or by changing the organisation. The latter rarely happens.

When developing data architecture, one must heed the old maxim, that all models are wrong, but some models are useful. Architects love to design intricate beautiful structures and thus it is easy for them to go overboard with much too complex or detailed models. The target audience of the architecture blueprint and their needs should always be kept at the forefront with things kept as simple as possible. The latter can become an important factor, as an architecture blueprint is a liability not an asset. It requires constant upkeep that the organisation must be willing to pay for (See 5.1.) and the more complex the model is, the more expensive is going to be to keep it up to date and in line with the actual physical reality. Therefore, it is always advisable to err on the side of less detail, when in doubt.

Data quality is best addressed as high upstream, as possible, preferably at the point of data creation or acquisition. This simplifies data lineage management as well as data management in general. If a business process creating the data does not yield high quality data, little can be done downstream to fix it. Also, the question of data quality can become intertwined with business logic of the service offered. For example, if a data subject reports their heartrate to be 688 beats per minute, this is bound to be a data quality issue – the question then becomes whether a later, seemingly more “realistic” data entry, should be interpreted by the solution as a signal about the user’s improved well-being or if the solution takes deviations and input errors into account.

In this, user interface design plays an important role: users need to understand, what data they need to give to generate high-quality data. If a user cannot find a non-compulsory field to fill in, they will not fill it; if they cannot find the button opening another section of the form, they will not fill it in.

For managing data quality, many sets of data quality attributes are suggested by various sources. In addition to picking the ones that best reflect what data quality means for your organisation, all attributes should also be accompanied by specific metrics to enable SMART goal definition. Table 2 contains some examples of data quality attributes and the metrics potentially associated with them.

ATTRIBUTE	SAMPLE METRICS	DESCRIPTION
Consistency	range, variance, standard deviation, fit to a distribution	How stable the data is, to what extent it behaves in a statistically predictable manner
Accuracy	error ratio, standard deviation, noise levels	How accurately the data reflects the physical reality
Completeness	% of completed records	To what extent one can expect the required data to be there
Auditability	% of altered data, % of untraceable data	To what extent the data lineage is under control
Validity	% of structurally valid records	To what extent data corresponds to the expected structure
Uniqueness	# or % of repeated records	To what extent the data describes unique objects or observations
Timeliness	mean age of records, time variance	To what extent the data reflects reality rather than the past

Table 2. Examples of data quality attributes and the corresponding metrics. Source: Authors.

Data-focused processes will have to tackle the issue of semantics, i.e., the meaning a given data element carries. A given body mass index can convey both wealth and unhealthy eating habits in different cultural context while "blood pressure" can mean systolic and diastolic pressure measured in different ways. This can make communication difficult both between and within organisations.

To alleviate the issue, numerous standards and ontologies (the HL7 family, UMLS, WHO-FIC etc.) exist to create a common meaning to data, especially between organisations and information systems. Because of the complexity of the field and the inherent vagueness of the concept of "meaning", these standards tend to be complicated and, while often useful, can easily become overwhelming and lead to differences in interpreting the standard itself. It is thus recommended organisations implement the minimal useful semantics-related processes moving forward only with tangible gains in sight. Due to the multitude of the standards and complexity of their implementation, any data sharing plan should also involve a process by which semantic interoperability is assured between the stakeholders.

5.2.2.3. Metadata management

Metadata management is a complex field because of its tautological nature: all data about data (including the results of the data tracing process) is also data and thus subject to the same managerial processes. This includes metadata creation about metadata. Thus, metadata management can be seen as implementation of the same data governance model with a different business value definition, strategy etc. Herein also lies a key recommendation on metadata management: assure it creates value and assure some of this value is directed back to developing metadata governance. The tangible benefits of metadata management should always exceed the expense of its creation and upkeep.

5.2.3. Tools and technology

The tools and technology an organisation uses for achieving and maintaining control over its data are very much dependent on:

- The data architecture and system architecture which, in turn, are dependent on the organisational architecture. For example, a siloed organisational architecture might require specific in-house data exchange platform to allow for different siloes to follow their own data management practices while allowing for the whole organisation to both execute its data strategy as well as fulfil its legal obligations.

- Data governance and management processes implemented. For medical applications, compliance processes usually have high importance and thus the required tooling needs to be very capable and potentially bespoke. For data analytics or brokerage organisation, data lineage and integration tooling are of crucial importance and thus need to be extensively developed. One should avoid implementing tools simply because the best practice says so without being able to clearly articulate the benefits delivered (See also the CAPEX vs. OPEX discussion in 0).

One of the key decisions to be made around tools and technology is a build vs. buy decision. Regardless of the circumstance, the combination of the two (acquire and modify extensively) is very rarely beneficial as it combines the downsides of both alternatives with little additional upside. In general, one should build tooling if the organisation has either superior understanding of the business process or is demonstrably capable of building a better tool for the job. In all other cases one should simply acquire the necessary tooling.

The second main decision on tools and technology is the question of lock-in. Although vendor lock-in is a known term, all technology decisions involve locking oneself into a given ecosystem of standards, practices, and stakeholders. For example, an open-source tool can be abandoned by its maintainer and prove costly to replace. Even following a well-known industry standard, like HL7/FHIR will make it difficult to move off it once the decision has had long enough to influence other decisions in the organisation. Thus, tools and technologies should be chosen based on which ecosystems surrounding them provide the best long-term value for the organisation.

5.3. Control

Summary: *The section provides practical guidance on implementing the concept of data control described previously. Special attention is given to metadata management and information security along with its sub-domains.*

Lead questions: *What risk management, mitigation and response mechanisms can be applied to your organisation? How do you assess and ensure upstream and downstream compliance, including standard data management requirements for procurement, contracts with vendors, data sharing agreements, cloud procurement, third-party development, and licencing transactions? What mechanisms can you implement to solidify your cybersecurity outlook and strengthen incident response?*

5.3.1. Risk management

Risk management is a wide and specialised field often appearing impenetrable to a bystander. Indeed, the following does not seek to provide a comprehensive set of guidance on risk management as, almost always, the best guidance is to consult a specialist. To do so effectively, knowledge of the structure of the field (depicted on Figure 6) is useful.

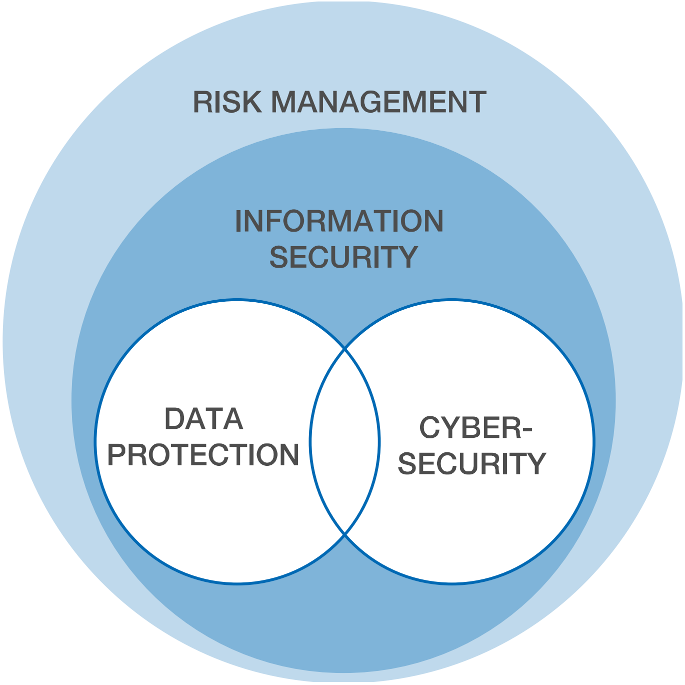


Figure 6. Fields of risk management. Source: authors.

Risk management, as a field can be divided up for the context of data governance:

- **Risk management** deals with all risks (including, for example, risks stemming from its physical environment) an organisation encounters. Risk is defined as the impact uncertainty has on goals of the organisation and is measured as a function of probability of the undesired event occurring and the impact of it happening.

- **Information security** deals with risks related to information. Based on the hierarchy presented on **Error! Reference source not found.**, this involves all risks related to data, but is wider in nature. Somebody determining the number of people in the building by counting people entering and leaving is an information security risk but does not involve data.

- **Data protection** is a field focused on risks around regulated data and is thus adjacent to the field of compliance management. Regulated data is often regulated, because it is of interest to malicious actors, thus the overlap with cybersecurity.

- **Cybersecurity** is focused on information security risks related to malicious actors. Somebody forgetting a laptop with patient data in a bus is a data protection but not cybersecurity incident because regulated data is involved without malicious intent. Somebody stealing said laptop from the office is a cybersecurity incident due to presence of malicious intent. An information system being taken down by a denial of service (DDOS) attack is a cybersecurity but not data protection incident because there was malicious intent, a detrimental effect on information security (more precisely, the availability attribute), and no specific targeting of regulated data.

As per the definition of risk, all risk management depends on the goals of the organisation being well defined. If there is only a vague knowledge of what we'd like to happen, there cannot be specific knowledge of what we do not want to happen. Also, it is important to have the risk appetite of an organisation clearly established to tell acceptable risks from unacceptable ones. Both assume certain maturity, size, and resources from the organisation. As incidents unfortunately do not care about how mature an organisation is, a gap is usually created, where systemic risk management is not yet implemented but there are significant information and data assets present to be protected. In the early stages of an organisation or in strategically challenging environments, it is recommended to focus on data protection and focus on all mandated risk mitigation measures. This provides a stopgap until the organisation is ready to take on comprehensive risk management and mitigates compliance risks from the other.

The risk management process consists of the following steps:

- **Context establishment,** where the goals of the organisation, its environment, assets, level of control over data required, perimeter etc. are analysed. Establishing boundaries of the organisation is a crucial part of context creation as these define the scope of risks to be managed. As an example, there is a vast difference between a policy A that does not allow laptops containing patient data to leave company premises and policy B that extends the boundaries to any system that contains company data, thus leading to regular employee trainings on why they should never leave their company phones unlocked.

- **Risk identification,** where risks are identified.

- **Risk analysis,** where risks are analysed from their probability and impact perspective. Risk relationships are also identified. Typically risks form a tree-like structure, where a single undesired event happening is caused by several risks materialising that in turn require several circumstances to arise.

- **Risk evaluation**, where risks are compared to the acceptable baseline and a priority list of risks to be tackled is created.

- **Risk treatment**, where risk treatment measures are put in place after choosing the risk treatment method from among **avoidance** (decide not to do the risky thing), **mitigation** (decide to put additional measures in place to reduce the probability or impact of a risk), **transfer** (decide to share the risk with somebody else by, for example, getting insurance) or **accept** (decide to do the risky thing anyway to get the benefits).

All steps described above are subject to continuous monitoring and review as well as communication and consultation with the members of the organisation. Regular monitoring ensures risk management to keep on top of inevitable changes in the organisation and its environment.

5.3.2. Information security

The section outlines main approaches to information security (ad-hoc, standards-based, architectural) and refers to main information security standards and frameworks relevant in the health sector. An emphasis is put on the idea of holistic information security: it is the organisation along with its hardware, software, and people, that is either secure or insecure rather than just software. Practical guidance on managing the three aspects of information security (availability, confidentiality, and integrity of data) is given.

Information security in the data governance context can be seen as managing three attributes of data:

- availability (can we access our data?),
- confidentiality (can unauthorised parties access our data?)
- integrity (has the data been changed?).

This implies the ability to measure these attributes for any given dataset as well as desired attribute levels having been set. Health data has commonly high requirements for integrity (changing somebody's blood type can be lethal) and confidentiality with availability varying depending on application area (secondary data analysis can be delayed without consequence while the ER team must have immediate access to allergy records).

The single most important practical understanding of information security is, that security is an emergent property of the entire organisation consisting of people, software and the physical infrastructure containing both. An emergent property is a property that only comes to light when individual elements are combined: both children and hammers are usually considered reasonably safe whereas their combination is not. Therefore, any approach to information security must consider all three elements which organisations are made of, to have a chance of success. For this reason, information security is a leadership challenge first, managerial challenge second and finally a technical challenge.

Secondly, nobody seeks to build insecure systems, neither software nor organisations. It is over time, as both organisations and their environment gradually change that their combination gradually becomes insecure. Therefore, any approach to information security must contain a mechanism for assuring things are secured continuously.

There are three main ways an organisation can tackle information security.
Firstly, information security can be based on an established standard. These standards vary in scope, methodology and application area but can allow an organisation to be sure all the necessary boxes have been ticked and prove that to be true via an audit process.

The most common information security standard is the ISO 27000 series of standards, where ISO 27001 covers information security management and ISO 27002 provides security techniques in terms of controls to be placed in the organisation. ISO 27799 applies the latter in healthcare. In addition, IEC 82304 provides requirements to manufacturers of health software products designed to operate on general computing platforms.

Besides ISO/IEC standards, various trade bodies and commercial entities have developed their own sets of standard operating procedures. Most notable of them is the CIS Cybersecurity Framework developed by SANS institute. Finally, many national security standards exist, that are either required in health sector or have special subsections dedicated to it. The two most notable are the Grundschutz and NIST, applicable in Germany and US respectively. Because of their comprehensiveness and versatility, these standards have been applied in other countries in part or in full or have been used as a basis for developing national standards. The main challenge is to adhere to the chosen standard without letting it turn into a formality.

Secondly, organisations can take a simple ad-hoc approach conducting risk analysis and implementing measures based on expert opinions, intelligence gathered and practical needs. The key downside of this approach is, that it is hard to tell, if and how sensible the processes and controls in place are. When implemented properly, ad-hoc information security can be more effective, than any standard implemented blindly. The key question here is, how can the organisation be sure, if theirs is done "properly". Implementation of this approach depends heavily on expertise of the information security personnel in place and the ability and willingness of the rest of the organisation to take their input seriously.

Thirdly, organisations can take the architectural approach focusing on designing organisational systems that are inherently safe. In this setting, the question is not "how can we make system X secure?" but "how do we design a system where security of system X matters the least?". One of the most widely known architecture-driven methods STAMP (referenced above) stems not from information security but from system safety. This method is centred around placing controls around potentially unsafe system components and then placing controls around *that* considering all the ways in which the controls can fail.

The organisational approach to information security can be applied in conjunction with other approaches as it can significantly ease their implementation by, for example, careful modularisation of the system in question. Implementing architectural information security implies a reasonably mature and reasonably skilled architecture capability.

5.3.2 1. Data protection

Data protection is, by definition, an area tightly linked to fulfilling responsibilities stemming from applicable regulation. The key regulations in the health sector in the EU context and their main requirements are described in 0 below. While the regulations contain their own principles and requirements, following a few key guiding points will make data protection easier and minimise risk:

- **Always have a legal basis for processing regulated data.** No piece of regulated data should touch the information systems of an organisation without it having a good legal reason to do so regardless of whether it stems from a contract, a consent given by the data subject or regulation. This allows all regulated data to be governed to have a clear set of requirements to which they must be compliant.

- **Ensure downstream compliance.** Whenever data moves between information systems or between organisations, the source of data should make sure the data recipient is equipped (both compliance-wise and practically) to take care of it. Within an organisation, this principle makes sure data lineage is not broken and that regulated data is perceived as such throughout its journey through the organisation. Between organisations this principle protects the data subjects (i.e., customers of the organisation) and allows the responsibilities of parties to be clearly formulated.

- **Assess upstream compliance.** Whenever receiving data from another information system or a third party, the legal basis of data processing, level of compliance etc. should be assessed by the receiver. This allows to reduce errors from misclassification of data and provides a good input for data governance processes including data tracing. However, it also allows for more accurate risk management by identifying incoming data that might have been collected without a proper legal basis, might have been unduly altered etc.

- **Make policies available to data subjects.** Documents guiding data governance in the organisation should be made public and easy to find for current or potential data subjects. This creates transparency underpinning trust between the organisation and the data subjects but also fosters accountability within the organisation.

5.3.2.2. Cybersecurity

Figure 7 depicts main cybersecurity terms and their relationships. The key message here is, that there is a long chain of events leading to a loss: a vulnerability creates a threat that an attacker can utilise to breach an organisation leading to a penetration that, via a compromise, can lead to a loss. From one hand, this means that when a chain can be broken at one point, loss will be prevented. For example, if we can deploy PR function in the right way, it can reduce the number of people willing to conduct an attack and subsequently reduce losses.

From the other hand, there are so many available combinations for each of the links in the chain starting from vulnerabilities, that a successful chain is always likely to be found. This is the reason there are no impenetrable systems. A determined attacker will always gain their prize.

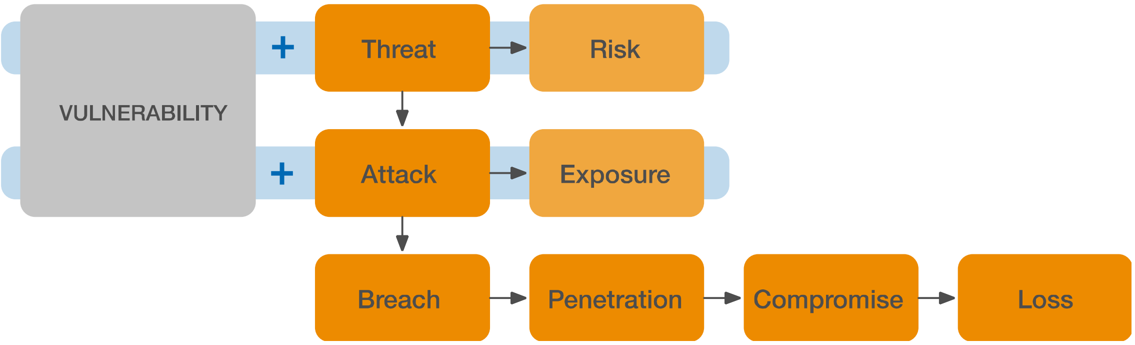


Figure 7. The chain of loss occurrence. Source: e-ITS Risk Management Guide, ISO/IEC 2382:2015

Observe that the links in the chain of loss occurrence are very different by nature, combining disciplines from cryptography and software engineering (vulnerabilities) to communication (either communicating oneself as a target to attackers or not doing so) to economics or politics (the reason a compromise is beneficial to an attacker). Cybersecurity is therefore a cross-disciplinary domain: all training, exercises, incident response protocols etc. should always encompass all the relevant parts of the organisation rather than just technical personnel. No spokesperson ever drawing public ire on the organisation or making statements about the organisation being impenetrable is a very useful cybersecurity measure.

Especially for smaller organisation, it is important to understand the extent to which each link in the chain of loss occurrence can be monitored and incidents responded to. The acceptable risk levels stated should be on par with what the organisation can afford. It is advisable to outsource some of the effort by joining local cybersecurity communities and contributing to joint exercises. As a part of a bigger community, the chances of learning about relevant emergent threats on time or being able to field a breach are much higher as part of a network of trusted specialists.

Although cybersecurity is a complex field, some measures to be taken are specific to protecting health data:

- **Minimisation.** As discussed previously, data is a liability. The less data there is, the less data must be governed and the smaller the risks. Not just data as a whole should generate more value than it takes to govern it, but the same is true for all data elements. Since risk is a significant source of cost, data minimisation is an important cybersecurity measure. It also makes sense to minimise not only the data itself but its usefulness to an attacker. If, for example, there is a need to cluster but not filter data by a post code, there is both no need to store the complete address and the post code can be obscured by a one-way function.

- **Anonymisation**. Anonymisation allows, in theory, to make health records anonymous by removing references to the data subject. Strictly speaking, there are two separate techniques: anonymisation, that removes all references to the data subject and pseudonymisation that replaces the data subject reference using a one-way function. The techniques can be applied if the data does not need to be linked back to its subjects e.g., for use in self-service systems. However, anonymisation does not prevent data from being re-identified by cross-referencing it with other data sources, applying machine learning techniques or using cluster analysis. Therefore, usefulness of anonymisation depends on how the acceptable risk level relates to the belief, that the captured data can be re-identified at any point in the future using any dataset available at that point.

- **Encryption**. Encryption makes data-at-rest (i.e., data that is statically stored) unreadable by anyone without the decryption key. Encryption is a useful technique to protect sensitive datasets but has two significant caveats. Firstly, data must usually be decrypted to be used. Hence, usefulness of encryption depends on the security of all systems that can read the decrypted data. Secondly, an attacker with access to the system can still extract the encrypted data and then proceed to attack its encryption at their leisure. These attempts are likely to succeed eventually, as all cryptography becomes weaker over time – not just because bad actors are continuously working to uncover weaknesses, but also because the computation power of hardware increases exponentially, thus making it more likely for brute force attacks to succeed. Hence, for encryption to be useful, usefulness of the encrypted data must decay faster than cryptography used**.**

- **Separation.** Separation can be seen as a combination of minimisation and anonymisation: data is split up between several individual systems in a manner that makes the data shards useful for their individual applications but requires the attacker to breach multiple systems to assemble a useful dataset. For example, instead of a central database containing a name, weight, and height of a data subject three separate individually protected databases can be envisioned with one containing names, the other weights and the third one heights of the data subjects. Statistical analysis is possible for both weights and heights of data subjects, but an attacker would need to breach all three systems to gain information on the body mass index of the data subjects.

5.4. Data Management

Summary: *Data management, especially in a complex environment like healthcare, is not an endeavour limited to a single organisation nor does data management happen in a vacuum. This chapter focuses on main interaction points between the data management structure described above and the surrounding context. In the context of active and healthy ageing, however, several specific infrastructure elements are frequently needed and are thus discussed below.*

***Lead questions:** What role can flexible and open data management play in your business model and external relations? How do you best set up data sharing mechanisms? What legal implications can the use of AI have for your organisation?*

All regulated data being processed should have a legal basis for doing so. One way to gain that legal basis is to ask the data subject for the consent. According to GDPR, the consent to process data should be given by a clear affirmative act of the user and, if there are several purposes for processing data, they all require a separate consent – see also (European Commission, 2016). To ensure all data items have such a consent attached, to keep track of the different types of consents through data lineage and to allow the data subject to revoke their consent, a consent management system (be it a human-based, paper-based, or digital one) is often necessary. The consent system becomes especially critical, if the organisation allows third parties to access data based on the consent of the data subject.

Such a consent management system can range from a simple database recording consent events to a full-blown consent workflow with complex third-party integrations. From the data governance perspective, the consent system should at the very least allow each consent issued to have a specific identifier, that can be processed along with the data item and provide the facility to confirm validity of a consent by that ID. As consent revocation is a relatively rare occurrence, oftentimes a reverse system is useful, where the consent system broadcasts a list of revoked consents so that data processing systems can act accordingly.

Often, data processors have either a commercial interest or an obligation to publish open data. Open data usually means data that is publicly available without undue barriers for access. This definition does not rule out authenticating counterparts accessing the data nor does it prevent data providers from issuing license terms along with the data. The following principles should be followed whenever publishing open data:

- Conduct a thorough impact analysis on the process of publishing a dataset for assessing the risk and proposing mitigation measures for cases where data can be re-identified at some point in the future, and only publish datasets with a suitably low risk level.

- Always provide a license by combining current best practice in your legal context with your requirements in terms of rights to change or re-publish the data, commercial use of data or the requirement to publish derivative datasets.

- Prefer API-based access to publishing static datasets to ease data re-use and assure freshness of the results.

- Publish thorough documentation on both the API (or file format) and the semantics of the data along with the dataset. Keep the documentation up to date.

- Publish thorough metadata of the dataset including data lineage.

- For open data sets requiring authentication, use standard JWT tokens. JSON Web Tokens are an open, industry standard RFC 7519 method for representing claims securely between two parties.

Sharing non-open data is a subset of sharing open data in a sense that it also involves sharing data with third parties but limits whom the data is shared with. The same principles apply, as for open data, but there are severe limitations caused by the need to firmly secure the data exchange.

Firstly, a standardised data exchange or interoperability platform might already exist in your context, be it a national interoperability platform, a healthcare data exchange, or a combination of both. Use of an existing platform is always preferable to figuring out and operating all the security details like certificate expiry, key rotation etc. Secondly, assure all data sharing has a solid legal basis and that the data recipient has the right to process the data received. This might be necessary to be stipulated in a contract between the parties (see 5.5.3).

> *To provide a seamless care journey, it is important that relevant technologies in the health and social care system are interoperable, in terms of hardware, software and the data contained within. For example, it is important that data from a patient's ambulatory blood glucose monitor can be downloaded onto an appropriate clinical system without being restricted to one type. Those technologies that need to interface within clinical record systems must also be interoperable. Application Programme Interfaces (APIs) should follow the Open API Best Practices, be documented and freely available and third parties should have reasonable access in order to integrate technologies.*
>
> *Good interoperability reduces expenditure, complexity and delivery times on local system integration projects by standardising technology and interface specifications and simplifying integration. It allows it to be replicated and scaled up and opens the market for innovation by defining the standards to develop upfront. (Source: NHS, Digital Assessment Criteria-DTAC)*

Artificial Intelligence (AI) is an important way in which data can be managed to create value. The acronym is sometimes also said to mean "Augmented Intelligence" to denote the fact, that the AI is merely an extension of human intelligence and not intelligence in itself. The key properties of AI to be mindful of are:

- AI, as a mathematical construct, is fundamentally impartial, but the training data it is fed, the target functions etc. can be biased either deliberately or by accident.

- AI is emergent in nature. Under the hood, it consists of a mathematical structure and a set of parameters continuously adjusted to give the "best" (for some meaning of the word) output for a given input. Thus, the results emerge when two separate sets of data are simultaneously fed into a mathematical structure. The system is usually complex enough for it to be at least somewhat unpredictable as to what the output of the system is going to be for a given input. Also, there is no algorithmic explanation as to why an incorrect result was given. For a regular computer program, the program logic can be analysed, and explanation devised, but for AI we mostly only know that the mathematics yielded an incorrect result for this input.

- From the legal perspective, AI requires caution. For example, the GDPR Article 22 requires data subjects not to be subjected to decision-making based solely on automated processing which "... produces legal effects concerning him or her or similarly significantly affects him or her" (European Commission, 2016).

5.5. Context management

Summary: *This section offers a brief excursion into surrounding aspects that affect your ability to implement your vision of data governance. These factors include the use of cloud infrastructure, the nature of different risks, and effects of various legal environments.*

Lead questions: *Do the benefits of cloud infrastructure outweigh the risks for your organisation and associated business models? To which legal frameworks will you have to pay particular attention to ensure compliance?*

5.5.1. Infrastructure

Although managing technical infrastructure is not part of data governance in this guidebook, managing data requires significant amounts of infrastructure to process and govern. Modern data management infrastructure is mostly cloud-based. In the rare case private infrastructure is justified, that too will be fundamentally based on cloud infrastructure.

The main benefit of public cloud services is, that they abstract and scale up the unsightly parts of infrastructure management. This can easily lead for the cloud customers to lose sight of the fact that there is no cloud, there is somebody else's computer others area allowed to use. This significantly alters the risk profile of infrastructure. Most of the risks are (usually) acceptable but need to be considered nevertheless:

- The customer of cloud computing is fundamentally not in full control of the data being processed using cloud infrastructure. Since the data is stored on the physical hard drives of the cloud service provider, their employees can look at the data, make copies of it, etc. This includes encrypted data, as all cryptography degrades over time. Organisations should therefore adjust their desired level of data control.

- The cloud infrastructure is shared with other tenants of unknown profile. Virtual machine escape vulnerabilities allow the attacker to gain access to the sensitive parts of the operating system running the virtual machine they have access to. This in turn can grant them full access to other virtual machines running on the same physical hardware or, in worst case, on any hardware operated by the cloud provider. Such risks should be considered and mitigated as part of the information security risk management process.

- Cloud infrastructure providers are legally complex multinational organisations subject to a bewildering array of rules and regulations. These regulations will inevitably conflict with each other and, therefore, each cloud provider has a mechanism to prioritise regulations to be compliant to. Basically, if regulation A states that nobody should be able to read a file and regulation B states somebody must be able to read a file and both regulations are equally applicable, there is no telling if somebody will be able to read the file or not.

  A customer to the cloud service provider has minimal control over and knowledge of these mechanisms. Moreover, because of the legal structure of the cloud service providers, customers do not even necessarily have full visibility on what jurisdiction might potentially apply in each context. Cloud customers can typically choose, which physical locations will store their data and have some control of the jurisdiction applied but further control is commonly not possible. That choice should therefore be deliberate and in line with legal obligations of the organisation.

- Cloud infrastructure providers are businesses with their own interests and lifecycles. They go bankrupt, get bought and sold, close non-profitable ventures etc. Organisations should therefore always seek to keep their infrastructure as cloud-neutral as possible and have a well-defined plan for switching service providers. Feasibility of a on-premises backup of all the critical content of the cloud should be considered as an option, as this would allow recovery of data in case of a catastrophic failure on the cloud provider side.

In addition to changes to risk profile of the organisation, use of cloud services creates a need for the process of cloud management. In case all members of the organisations are just allowed to spool up virtual machines to simply be abandoned, costs will rapidly spiral out of control. Instead, the cloud infrastructure should be carefully designed to match the needs of the organisation and continuously managed in terms of costs and structure. This includes both low-level tasks like configuring and securing network access to the cloud and high-level advisory tasks like evaluating value-added services of the cloud provider and advising members of the organisation on their use.

Cloud security operations are an important part of this process. The fact, that an attacker does not know precisely where to look for a database with no or default security controls, does not constitute a security measure. Because of many interesting incidents, resources of all major cloud providers are by now regularly probed for misconfigured databases by automated process run both by security researchers and malicious actors.

In this regard, two common failure points are the failure to secure access to cloud resources leading to sensitive databases being left world-readable and the failure to properly configure the credentials used to configure the cloud services provided. The former will lead to data leaks and the latter to potential loss of the entire infrastructure if the single account used to configure it is compromised, or the person leaves the organisation etc.

5.5.2. Risk and security

As discussed previously, risk management is an integral part of data management. However, data management can cause risks to the organisation itself. These need to be managed by the general risk processes of the organisation but have their roots in data governance and thus have an impact on it. Typical examples of categories of such risks are as follows:

- **Risks stemming from low quality of data**. Such risks expose an organisation to adverse effects because the data they process is of insufficient quality. This can either be caused by the data quality management process failing or incorrect input being given to it about the required level of quality. (The STAMP model provides a methodology for addressing failure modes of sub-processes.) There are limits to the ability of the data quality process to detect incorrect data.

  For example, if incorrect blood group data (an accuracy quality attribute) is provided by the organisation by a third party or the data subject themselves and then used by the organisation with fatal consequences, there is little data governance processes could have done about it. Such risks therefore require organisation-level mitigation.

- **Risks stemming from mistakes in data processing**. These risks are caused by data management processes failing in some fashion. A typical risk in this category is the risk caused by AI misclassification. As such systems do not guarantee full accuracy of the results, they can produce incorrect results that can cause risk. Also, algorithmic data processing can contain errors due to, for example, either software quality assurance or software analysis processes failing.

- **Reputation risks.** These risks are caused by the nature of data processing being perceived as undesired by the stakeholders. Especially in the field of health data, questions of ethics frequently rise that are not necessarily obvious on the level of individual business processes within data governance. Also, combining various fully compliant and ethical processes together can yield results that can be perceived as unethical, biased or undesirable in other ways. Finally, these risks can be caused by failing to communicate complex data management processes correctly.

5.5.3. Legal and organisational context

The list of external stakeholders in the health data ecosystem is not static and is dependent on the legal and regulatory landscape below (see 0). There are entities actively taking part in the data processing (e.g., controller or processor) and there are those who do not process data, but set rules and guidelines how such data should be governed and processed through guidelines and supervision (e.g., National Health Authority or Data Protection Authority).

One key piece of legislation in managing and processing personal data is surely the GDPR which sets out four key roles that can be filled by an organisation or company besides the data subject and supervisory authorities (European Commission, 2016).

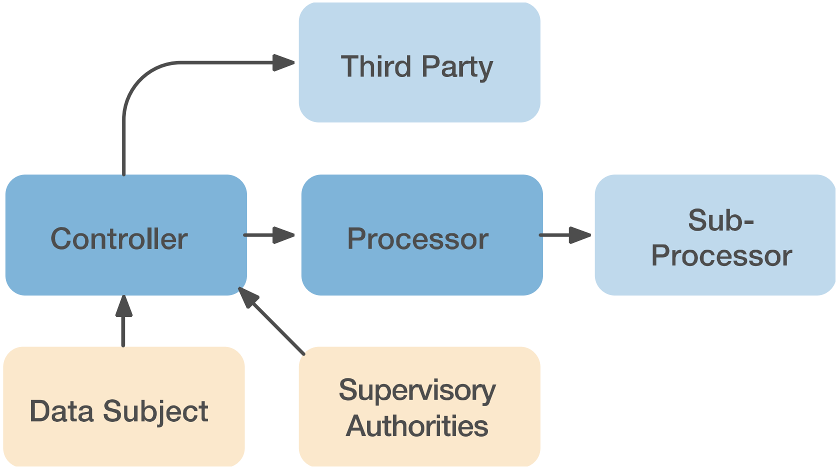


Figure 8: Stakeholders of GDPR. Source: Authors.

The controller is defined as any natural or legal person, public authority, agency, or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data (GDPR, art 4(7). Therefore, any company that determines why and how data processing activities are undertaken is considered to be a controller. The controller may use one or more processors who will process the data on behalf of the controller in line with GDPR, Art 4(8). The processor may in turn use sub-processors who process the data on behalf of the processor if the use of sub-processors is permitted by the controller. Art 28 of the GDPR sets out the requirements for formalising the relationship between the controller and processor and the use of sub-processors.

Such agreements are called data processing agreements or data sharing agreements and are crucial for: 1) helping all the parties be clear about their roles; 2) establishing the purpose of the data sharing/processing; and 3) covering what happens to the data at each stage.

Controllers and processors not established in the EU may be represented in the EU by a representative. In case personal data is made available to a recipient other than the controller, processor, or data subject (i.e., a third party according to GDPR Art 4(10), it is important to map them as relevant stakeholders. According to recital 54 of the GDPR, third parties in the context of health data processing could be employers or insurance and banking companies.

The two roles represented in white in **Error! Reference source not found.** are the data subject and supervisory authority, both of whom can become (sub-)processors of personal data in relation to the data subject's requests, claims, or supervisory proceedings. They are depicted as such because they cannot be filled in by an organisation or company. It must be noted that there are differences in the level of enforcement by supervisory authorities in different Member States and the controller must always be aware of the competent supervisory authority in its specific jurisdiction.

Data subjects act as a key source of personal data and the processing of personal data between organisations and companies and data subjects is often but by no means exclusively based on consent of the data subject (see GDPR Art (6)(1)). The supervisory authorities of Member States act as the gatekeepers to see that the requirements of the GDPR have been met by the controllers who are responsible for meeting GDPR requirements under GDPR Art 5(2). Although the practices and interpretations of the GDPR are heavily influenced by the European Data Protection Board, it is not listed as the external stakeholder because it is a policy body and does not get involved in the data management ecosystem beyond policy influence.

Although the roles and stakeholders of the GDPR as set out in **Error! Reference source not found.** could be used outside the context of processing personal data, the list of stakeholders does not grasp the full complexity of the external ecosystem of stakeholders from other relevant regulations such as the Medical Device Regulation (MDR) (European Commission, 2017a) / In Vitro Diagnostic Medical Devices Regulation (IVDR) (European Commission, 2017b) or standards such as the ISO standard ISO/TS 82304-2:2021 on health and wellness apps.

A mapping exercise of relevant legislation in the AHA field identified that there were 12 roles defined based on existing or upcoming legislation under which a company or organisation may fall arising from the MDR, the ISO standard on health and wellness apps, the GDPR and the Data Governance Act (European Commission, 2020). The roles depicted in blue in 9 are those that cannot be filled in by companies, but public institutions and are relevant for every organisation or company in the ecosystem.

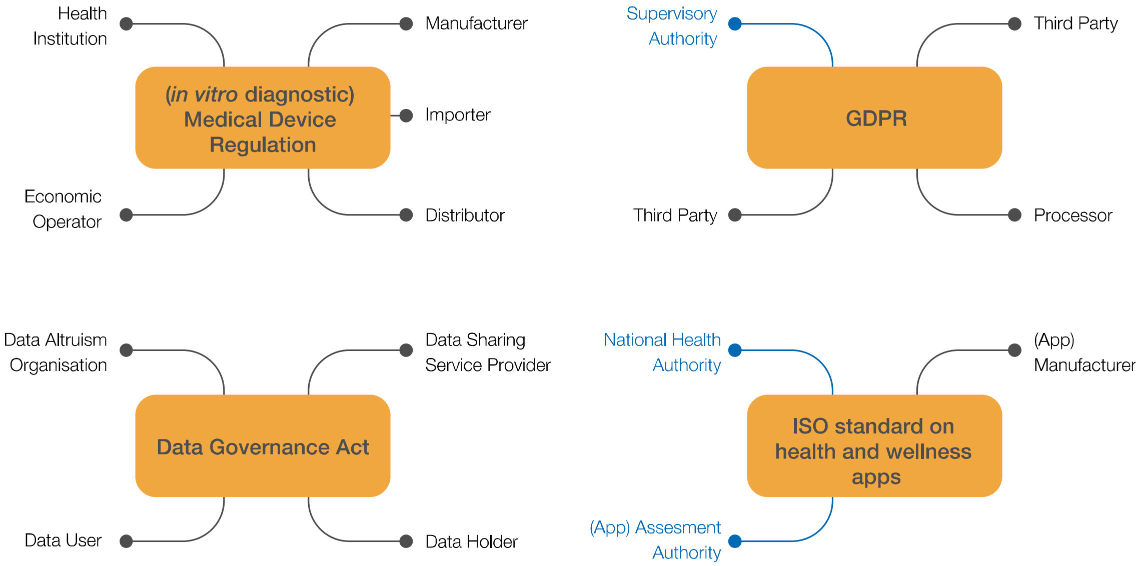

Health Institution
Manufacturer
(in vitro diagnostic) Medical Device Regulation
Importer
Economic Operator
Distributor
Supervisory Authority
Third Party
GDPR
Third Party
Processor
Data Altruism Organisation
Data Sharing Service Provider
Data Governance Act
Data User
Data Holder
National Health Authority
(App) Manufacturer
ISO standard on health and wellness apps
(App) Assesment Authority

Figure 9: Stakeholder roles from relevant regulations. Source: Authors.

Although the roles are separate from different legal acts, there could be overlaps – for example, the organisation acting as the controller within the meaning of the GDPR may also act as the App Manufacturer in the light of ISO standard on health and wellness apps and a manufacturer within the meaning of the MDR/IVDR. There is no exhaustive list of who may act as which stakeholder as the roles in essence are described broadly enabling any organisation or legal person to act as a given stakeholder if the requirements of the regulation are met (e.g., a company who defines the purposes and means of the data processing (i.e., controller) or it markets a medical device under its trademark (i.e., manufacturer).

5.5.4. Legal frameworks

This section outlines key EU level regulations in the health domain. Several of these were already mentioned in 0. This section provides further background on these regulations and adds other relevant regulations in the domain. Any digital health solution is subject to a wide range of regulations, from such as the GDPR, MDR/IVDR and more recent legislation such as the Data Governance Act and the proposed Data Act. However, digital health solutions also have to comply with a wide range of health sector-specific legislation at national and regional level, which can often impact on the way in which EU law is applied in practice.

General Data Protection Regulation

The GDPR sets out rules and requirements for controllers and processors when processing personal data of EU data subjects. The GDPR sets out the framework while specific legislation of Member States may still apply in the area of health and care (see below). The principle of accountability of GDPR Art 5(2) states that the controller is responsible for compliance for GDPR and must be able to demonstrate compliance at any given moment.

The GDPR introduced new definitions of “data concerning health”, “genetic data” and “biometric data” (see GDPR, Art 9(1)). With regard to the degree of their sensitivity and thus the need for special protection, (sensitive) data concerning health may encompass the subsets of the (even more sensitive) biometric and genetic data (TEHDAS, 2021).

Healthcare-specific GDPR-related safeguards include for example informed consent, pseudonymisation/anonymisation/de-identification, encryption, research ethics committee approval, technical and organisational measures for ensuring compliance with the GDPR. The GDPR safeguards should be integrated with other regulatory safeguards, provided e.g., by competition law, medicines regulatory requirements or ethical guidelines, cybersecurity requirements or the coming EU Regulation on AI (TEHDAS, 2021).

Medical Device Regulations – MDR and IVDR

MDR or IVDR applicability must always be considered when a (*in vitro* diagnostic) medical device is put to the market by a manufacturer or a distributor (European Commission, 2017). The definition of a (*in vitro* diagnostic) medical device includes a wide array of products (incl. software) which may be used to diagnose, monitor, prevent or treat people. Whenever your organisation or company processes health data and provides a service to end-users, it must be done in line with the MDR/IVDR. MDR/IVDR compliance requires certification of the (*in vitro* diagnostic) medical device and acquiring of a CE marking.

Data Governance Act

The Data Governance Act aims to create mechanisms for the re-use of public sector data that is conditional on the respect of the rights of others (notably on grounds of protection of personal data, but also protection of intellectual property rights and commercial confidentiality) and provide market rules for data sharing service providers who act as data intermediaries. The Act aims to encourage data availability and data sharing across EU and sectors.

Proposal for European Health Data Space

The EHDS is a health-specific data initiative which comprises rules, common standards and practices, infrastructures and a governance framework for the use and reuse of electronic health data. The EHDS proposal states its goals to promote optimal use of health data for healthcare delivery (primary) purposes as well as re-use for research and innovation, policy-making and regulatory activities. It targets the domain of digital health, covering health services and products, including tele-health, tele-monitoring and mobile health and proposes policies to enhance the development, deployment and application of trustworthy digital health products and services.

The initiative aims to improve the availability and quality of data in the healthcare sector, leading to fewer errors, less duplication of efforts and better medical outcomes The regulation seeks to standardise patient health files and ensure that electronic health data is interoperable and can be accessed across the union. Requirements would be introduced for Electronic Health Record (EHR) systems, the software used for storage and sharing of health records which must meet security and interoperability requirements.

A distinction is made between the usage of electronic health data for medical reasons (‘primary use’) and the re-usage of health data for activities such as research, medical algorithm training or policymaking (‘secondary use’). (European Commission, 2022).

ISO standard on health and wellness apps (ISO/TS 82304-2:2021)

The standard is relevant for all app manufacturers who wish to make and put to the market a health app as it describes the quality and reliability requirements. It is also relevant for app assessment authorities (e.g., ORCHA in the UK or Federal Institute for Drugs and Medical Devices in Germany) who evaluates health apps by applying their own methodology for assessment.

Member State Legislation

Art 9(1) of the GDPR defines health data as a special category of personal data. Article 9(2) defines possible legal grounds under which processing of health data may be permitted under Member State laws - processing is necessary for:

- reasons of substantial public interest (Art 9(2)(g)),
- the purposes of preventive or occupational medicine, for the assessment of the working capacity of the employee, medical diagnosis, the provision of health or social care or treatment or the management of health or social care systems and services (Art 9(2)(h)),
- for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices (Art 9(2)(i)),
- for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes (Art 9(2)(j)).

Please note that all the aforementioned possible legal grounds require additional legal basis from the relevant Member State's law.

The processing of health data may also be permitted if the data subject has given explicit consent to the processing (Art 9(2)(a)), except where Member State's law provides that such consent would not be sufficient for processing the health data. While the GDPR is not intended to apply to the personal data of deceased persons, Member States may provide for rules regarding the processing of personal data of deceased persons.

Article 9(4) of the GDPR further states that Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health. National legislations often define the responsibilities of service providers/data controllers about the usage of data, and they describe for which purpose it can be processed and what kind of limitations there exist, such as for which purpose the data can be used and in which format.

All this means that the bulk of legislation in the health domain is driven by Member State laws (European Commission, 2021).

References

Box, George EP. "Science and statistics." *Journal of the American Statistical Association* 71, no. 356 (1976): 791-799.

Conway, M. E., (1968). How do committees invent. Datamation 14, no. 4: 28-31.

DAMA International (2009). The DAMA guide to the data management body of knowledge. New Jersey, Technics Publications, LLC.

Data Processing Agreement Template. GDPR.eu. Retrieved from https://gdpr.eu/data-processing-agreement/, 25 May 2022.

Data Sharing Agreements. Information Commissioner's Office, UK. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/ico-codes-of-practice/data-sharing-a-code-of-practice/data-sharing-agreements/, 25 May 2022.

EIT Health (2021). Think Tank. Learning from health data use cases. Real-world challenges and enablers to the creation of the European Health Data Space.

European Commission (2022). Proposal for a regulation: the European Health Data Space. Retrieved from https://ec.europa.eu/health/publications/proposal-regulation-european-health-data-space_en, 3 May, 2022.

European Commission (2021). Assessment of the EU Member States' rules on health data in the light of GDPR.

European Commission (2020). Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), COM/2020/767 final. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767, 14 January, 2022.

European Commission (2017a). Regulation (EU) 2017/745 on medical devices (MDR). amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, OJ L 117, 5.5.2017, p. 1–175. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745, 3 March, 2022.

European Commission (2017b). Regulation (EU) 2017/746 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU, OJ L 117, 5.5.2017, p. 176–332. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745, 3 March, 2022.

European Commission (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj, 3 March, 2022.

European Data Protection Board (2020). Guidelines 05/2020 on consent under Regulation 2016/679, Retrieved from https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf, 3 May 2022

Gadiesh, O., Gilbert, J.L. (2001). Transforming corner-office strategy into frontline action. *Harvard Business Review* 79, no. 5

Graves, Desmond, ed. Management research: A cross-cultural perspective. Jossey-Bass, 1973.

ISO standard TS/82304–2, Health Software, Part 2: Health and Wellness Apps – quality and reliability, CEN/ISO, 2021.

ISO standard 13485:2016, https://www.iso.org/standard/59752.html

ISO standard ISO14971:2019, https://www.iso.org/standard/72704.html

ISO Guide 73:2009, https://www.iso.org/standard/44651.html

ISO/IEC standard 2382:2015, https://www.iso.org/standard/63598.html

ISO/IEC standard 27005:2011, https://www.iso.org/standard/56742.html

Leveson, N.G. (2016). *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

National Health Service. DTAC (Digital Technology Assessment Criteria). Retrieved from https://www.nhsx.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/, 12 January 2022.

OECD (2015). Health Data Governance: Privacy, Monitoring and Research - Policy Brief. Paris: OECD Publishing.

OpenEHR Foundation (2022). Open industry specifications, models and software for e-health. Retrieved from https://www.openehr.org/about/what_is_openehr, 30 April, 2022.

Phillips, J. J. (2005). *Investing in your company's human capital: Strategies to avoid spending too little--or too much*. Amacom Books.

Robbins, S. & Judge, T. (2012). Essentials of Organisational Behavior (11th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Silver, G. A., Silver, J. B. (1973). Data processing for business. Houghton Mifflin Harcourt Publishing.

The Open Group Architecture Framework 10th Edition. Open Group. Retrieved from https://www.opengroup.org/togaf/10thedition, 25 May, 2022.

Teece, D. J., Pisano, G., Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic management journal* 18, no. 7 (1997): 509-533

TEHDAS (2021a). Why health is a special case for data governance. Retrieved from https://tehdas.eu/results/, 3 March, 2022.

Wilkinson, M.D. et al (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 3/ 160018

World Health Organization WHO (2009). Practical guidance for scaling up health service Innovations. Retrieved from https://apps.who.int/iris/handle/10665/44180, 3 December 2021.

ANNEXES

Annex 1. Data management self-assessment checklist

There are many ways to evaluate if the data management processes in your organisation are functioning well and are adequate for the purpose (the value that they aim to produce). The basic evaluation should be done internally the organisation itself. This exercise will enable to map the data management functions as-is and give ground for making decisions, be it for allocating resources, re-organising roles and responsibilities, enhancing skills or implementing additional safeguards.

The questions below reflect the main topics covered in this guidebook. They draw input from a selected list of assessment and accreditation tools that are available for national use, including DTAC in UK, DIGA in Germany, and methodology used by Estonian Health Insurance Board to evaluate innovative health applications.

Privacy, data protection and consent

1. Does your solution collect personal data and/or sensitive personal data?
    a. If yes, do you have the means to confirm legally binding consent? (is it expressed as GDPR requires?)
    b. Can users withdraw their consent at any time?
2. Does your solution have a privacy policy and a notification informing the user of it?
3. Is the privacy policy accessible to the user during account creation and usage of the solution?
4. Does the privacy policy comply with all applicable laws, both national and international?
5. Does the user have the ability to view the data they generated?
6. Does the user have the ability to delete the data they generated?
7. Do the collaboration contracts with your partners stipulate the lawful storage and treatment of users' personal and non-personal data?

Data security

1. Do you use a GDPR-compliant storage solution? (on-premise, cloud, if so in which country)?
2. Do you have a nominated Data Protection Officer?
3. Do you have a response procedure in place to inform users and authorities of security incidents?
4. Do you use adequate encryption methods and channels to transmit all personal data?

Data usage, semantics and portability

1. Does the user have the ability to extract the data they generated?
2. Does your solution use standardised terminology (e.g. ICD-10 or HPO) and standardised clinical data modelling tools (e.g. OpenEHR) where applicable?

Data quality

1. Is the metrics for measuring data quality defined, as relevant to your business logic?
2. Do you have a specialist responsible for maintaining data quality?
3. Is there a quality maintenance process for the health information submitted by the user?
4. Do you practice data minimisation?
5. Is an appropriate retention policy established to erase or review the data stored?

Annex 2. Consent template

Consent is defined as one of the legal grounds of processing personal data under GDPR Art 6(1) and 9(2). Consent is mainly used by private sector entities when processing personal data. Therefore, it is crucial that the consent acquired from data subjects meets GDPR standards.

Consent is defined in Article 4(11) of the GDPR. Consent of the data subject means any *freely given*, *specific*, *informed* and *unambiguous* indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.

Therefore, the elements for a valid consent are:

- **Freely given**: means that providing the consent must be a real choice for data subjects. If the data subject has no real choice, feels compelled to consent or will endure negative consequences if they do not consent, then consent will not be valid as it is not freely given;
- **Specific:** means that the purpose of the processing must be specified as a safeguard against function creep; in case of multiple purposes, the controller or processor should provide granularity in consent requests (opt-in for each purpose to allow control);
- **Informed**: means the consent must include minimum requirements of information about the controller's identity, purpose of processing, what (type) of data collected and the right to withdraw consent;
- **Unambiguous indication of wishes:** means consent must always be given through an active motion or declaration.

The European Data Protection Board has given extensive guidance on the rules on consent in its Guidelines 05/2020 on consent (European Data Board, 2020).

The data subject must always be able to withdraw his or her consent. This means that controllers must be prepared to respect that choice and stop that part of the processing if an individual withdraws consent. This means the controller is not allowed to switch from the legal basis consent to legitimate interest once the data subject withdraws his consent. This applies even if a valid legitimate interest existed initially. Therefore, consent should always be chosen as a last option for processing personal data.

Consent may be acquired for a specific activity during a business process, or it could be defined in a Privacy Policy outlining the data protection practices of an organisation. Note that an offering with multiple hardware and software providers would require to ascertain a patient's consent individually for each provider in order for them to start sharing data with each other.

Sample: Our Company Privacy Policy

Source: https://gdpr-info.eu/issues/consent/

Our Company is part of the Our Company Group which includes Our Company International and Our Company Direct. This privacy policy will explain how our organisation uses the personal data we collect from you when you use our website.

Topics:

- What data do we collect?
- How do we collect your data?
- How will we use your data?
- How do we store your data?
- How will we use your personal data for marketing purposes?
- What are your data protection rights?
- What are cookies?
- How do we use cookies?
- What types of cookies do we use?
- How to manage your cookies
- Privacy policies of other websites
- Changes to our privacy policy
- How to contact us
- How to contact the appropriate authorities

What data do we collect?

Our Company collects the following data:

- Personal identification information (Name, email address, phone number, etc.)
- [Add any other data your company collects]

How do we collect your data?

You directly provide Our Company with most of the data we collect. We collect data and process data when you:

- Register online or place an order for any of our products or services.

- Voluntarily complete a customer survey or provide feedback on any of our message boards or via email.
- Use or view our website via your browser's cookies.
- [Add any other ways your company collects data]

Our Company may also receive your data indirectly from the following sources:

- [Add any indirect source of data your company has]

How will we use your data?

Our Company collects your data so that we can:

- Process your order and manage your account.
- Email you with special offers on other products and services we think you might like.
- [Add how else your company uses data]

If you agree, Our Company will share your data with our partner companies so that they may offer you their products and services.

- [List organisations that will receive data]

When Our Company processes your order, it may send your data to, and also use the resulting information from, credit reference agencies to prevent fraudulent purchases.

How do we store your data?

Our Company securely stores your data at [enter the location and describe security precautions taken].

Our Company will keep your [enter type of data] for [enter time period]. Once this time period has expired, we will delete your data by [enter how you delete users' data].

Marketing

Our Company would like to send you information about products and services of ours that we think you might like, as well as those of our partner companies.

- [List organisations that will receive data]

If you have agreed to receive marketing, you may always opt out at a later date.

You have the right at any time to stop Our Company from contacting you for marketing purposes or giving your data to other members of the Our Company Group.

If you no longer wish to be contacted for marketing purposes, please click here.

What are your data protection rights?

Our Company would like to make sure you are fully aware of all of your data protection rights. Every user is entitled to the following:

The right to access – You have the right to request Our Company for copies of your personal data. We may charge you a small fee for this service.

The right to rectification – You have the right to request that Our Company correct any information you believe is inaccurate. You also have the right to request Our Company to complete the information you believe is incomplete.

The right to erasure – You have the right to request that Our Company erase your personal data, under certain conditions.

The right to restrict processing – You have the right to request that Our Company restrict the processing of your personal data, under certain conditions.

The right to object to processing – You have the right to object to Our Company's processing of your personal data, under certain conditions.

The right to data portability – You have the right to request that Our Company transfer the data that we have collected to another organisation, or directly to you, under certain conditions.

If you make a request, we have one month to respond to you. If you would like to exercise any of these rights, please contact us at our email:

Call us at:

Or write to us:

Cookies

Cookies are text files placed on your computer to collect standard Internet log information and visitor behaviour information. When you visit our websites, we may collect information from you automatically through cookies or similar technology

For further information, visit allaboutcookies.org.

How do we use cookies?

Our Company uses cookies in a range of ways to improve your experience on our website, including:

- Keeping you signed in
- Understanding how you use our website
- [Add any uses your company has for cookies]

What types of cookies do we use?

There are a number of different types of cookies, however, our website uses:

- Functionality – Our Company uses these cookies so that we recognise you on our website and remember your previously selected preferences. These could include what language you prefer and location you are in. A mix of first-party and third-party cookies are used.
- Advertising – Our Company uses these cookies to collect information about your visit to our website, the content you viewed, the links you followed and information about your browser, device, and your IP address. Our Company sometimes shares some limited aspects of this data with third parties for advertising purposes. We may also share online data collected through cookies with our advertising partners. This means that when you visit another website, you may be shown advertising based on your browsing patterns on our website.
- [Add any other types of cookies your company uses]

How to manage cookies

You can set your browser not to accept cookies, and the above website tells you how to remove cookies from your browser. However, in a few cases, some of our website features may not function as a result.

Privacy policies of other websites

The Our Company website contains links to other websites. Our privacy policy applies only to our website, so if you click on a link to another website, you should read their privacy policy.

Changes to our privacy policy

Our Company keeps its privacy policy under regular review and places any updates on this web page. This privacy policy was last updated on 9 January 2019.

How to contact us

If you have any questions about Our Company's privacy policy, the data we hold on you, or you would like to exercise one of your data protection rights, please do not hesitate to contact us.

Email us at:

Call us:

Or write to us at:

How to contact the appropriate authority

Should you wish to report a complaint or if you feel that Our Company has not addressed your concern in a satisfactory manner, you may contact the Information Commissioner's Office.

Email/Address

Annex 3. Checklist for data protection impact assessment

Source: Information Commissioner´s Office, UK
https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/

1. **Describe the nature of the processing:** how will you collect, use, store and delete data? What is the source of the data? Will you be sharing data with anyone? You might find it useful to refer to a flow diagram or other way of describing data flows. What types of processing identified as likely high risk are involved?
2. **Describe the scope of the processing:** what is the nature of the data, and does it include special category data? How much data will you be collecting and using? How often? How long will you keep it? How many individuals are affected? What geographical area does it cover?
3. **Describe the context of the processing**: what is the nature of your relationship with the individuals? How much control will they have? Would they expect you to use their data in this way? Do they include children or other vulnerable groups? Are there prior concerns over this type of processing or security flaws? Is it novel in any way? What is the current state of technology in this area? Are there any current issues of public concern that you should factor in? Are you signed up to any approved code of conduct or certification scheme (once any have been approved)?
4. **Describe the purposes of the processing:** what do you want to achieve? What is the intended effect on individuals? What are the benefits of the processing – for you, and more broadly?
5. **Describe compliance and proportionality measures**, in particular: what is your lawful basis for processing? Does the processing actually achieve your purpose? Is there another way to achieve the same outcome? How will you prevent function creep? How will you ensure data quality and data minimisation? What information will you give individuals? How will you help to support their rights? What measures do you take to ensure processors comply? How do you safeguard any international transfers?
6. **Identify and assess the risks:** describe source of risk and nature of potential impact on individuals. Include associated compliance and corporate risks as necessary.
7. **Identify measures to reduce risks.**
8. **Record the outcomes of assessment.**