imec
embracing a better life

WHITE PAPER

Smart data operations for better data-informed decision-making

Contributors
Stefan Lefever, imec-EDiT
Philippe Michiels, imec-EDiT
Dimitri Schuurman, imec-EDiT
Tanguy Coenen, imec-EDiT
Nick Vintila, imec-EDiT

Version
January 31, 2023

Contents

About this white paper

Smart data in the context of decision-making is a widely discussed and researched topic, as demonstrated by the more than 10,000 hits on Google Scholar, with over 2,000 results from 2022[1]. This high level of interest has fostered many research and innovation projects in Flanders (Belgium) aiming to generate best practices based on local implementations and a variety of ecosystem-driven approaches. Drawing on learning from these various Flemish projects, this paper aims to demystify the 'behind the scenes' work that is needed to process and prepare data which can enable data-informed decision-making in the context of societal challenges.

This paper is intended for technical strategists, such as information architects and others, who work regularly with data and Information Technology (IT) to support better government and more effective responses to growing societal challenges – climate change, increased population density, public health concerns, safety and security, the use of scarce resources, etc. The paper aims to provide an overview of how data can be turned into meaningful information for decision-making support. It discusses the processes involved including the appropriate starting points and the importance of sketching out the opportunities and challenges that may arise. The discussion also highlights the importance of understanding the meaning of the available data and where and how to use it.

The theoretical discussion is complemented by real-world examples from various Flemish and international innovation projects. These include VLOCA (Vlaamse Open City Architectuur) and VSDS (Vlaamse Smart Data Space) and European initiatives such as Gaia-X, IDSA (International Data Spaces Association) and DSSC (Data Spaces Support Centre). These examples offer illustrations of how data can be made smarter for decision-making.

We warmly invite interested readers to join the open ecosystem of these initiatives and to contribute to the community and to the ambitious goal of achieving datafication [1] at societal scale.

[1] Results for search terms "smart data" & "decision making" on google.scholar.com as of 19/12/2022

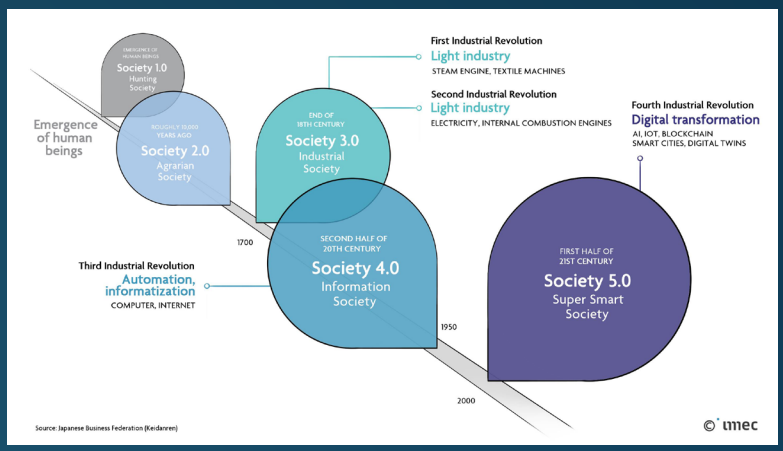
1. Introduction

Data-informed decision-making has gained a lot of momentum in recent times. Most newly installed IT processes, systems and applications generate data, or offer the ability to produce large quantities of data which can be used to enable more and faster insights for decision-making. Recent technologies and applications such as AI (Artificial Intelligence), the Internet of Things (IoT), blockchain, big data and urban (local) digital twins [2] allow for a plethora of new insights to be collected from the data produced (as in the concept of 'Society 5.0' [3]).

However, the ability to produce data is not sufficient on its own. 'Raw data' does not generate any comprehensible meaning. Hence, the importance of 'smart data' [4], which refers to both the output of a process which transforms raw data to a point where it is ready to deliver meaningful insights and to the process itself. In this paper we discuss the various steps and elements involved in the process of transforming raw data to generate meaningful insights that can be applied in decision-making. The learning presented is derived from practical experiences in a wide variety of Flemish innovation projects. These projects illustrate the process of generating smart data in a wide variety of ecosystems. Based on the lessons learnt, this paper seeks to offer practical guidelines for 'smart decision-making'. The colored 'illustration' boxes throughout the paper connect the thinking behind these specific examples with broader trends and phenomena in the domain of data processing and operations.

The rise of Society 5.0 [3].

"A Human-centered society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space"



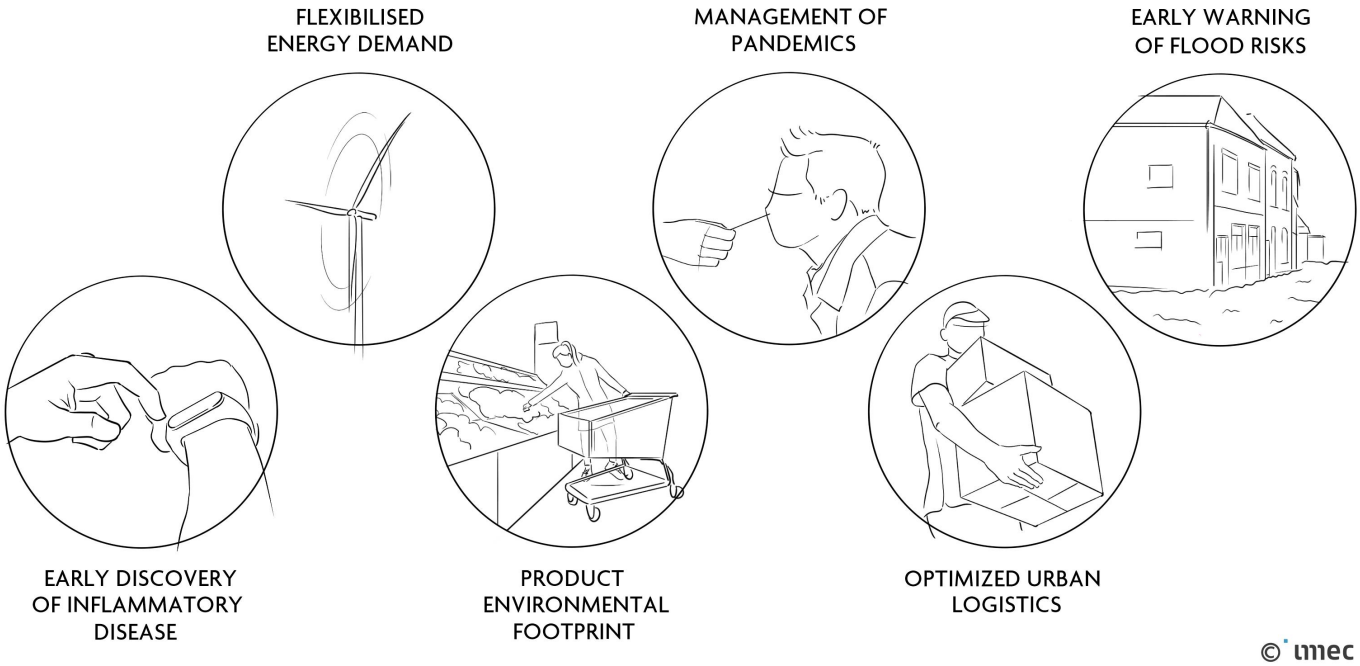
The various sections in the paper focus on particular aspects of smart data for decision-making. Section 2 sketches out the promising role for the use of data in addressing the exponential rise in complexity of societal challenges. Section 3 introduces how applying the FAIR [5] principles and techniques is a first step in improving access to, and reuse of, data at scale in specific locations and / or applications. It discusses why data intended for use in decision-making needs to be managed and governed. Section 4 highlights key challenges in making data more convenient for decision-making. Section 5 defines a second step in applying smart data principles, and how this is fundamental to managing and governing data in a decision-support context. Section 6 highlights some of the tools that support the goals of using smart data in decision-making, particularly those being developed in the Flemish ecosystem. And finally, Section 7 lists some takeaways and next steps.

2. Societal challenges and the role of data

Figure 1 illustrates several major societal challenges. Efficient and trustworthy collection, processing, and analysis of data (real-time/live-time/historical) can be a game / life-changer in addressing these challenges. For instance, data from smartwatches and fitness trackers can measure numerous health parameters continuously and feed doctors with crucial information. Given this capability, a question arises – what if it were possible to collect data from the entire population (while preserving individuals' privacy) in ways that enabled assessment of correlations between health, stress, pollution, pandemics, etc.? And what if it were then possible to use this information in preventative healthcare?

There are many other areas where such collection, processing and analysis of data has similarly significant potential to address ever-more complex societal challenges. One of these is matching energy supply and demand – large quantities of (real-time) data are now becoming available from smart meters and smart appliances which could enable balancing of supply and demand more intelligently, providing guidance for producers and consumers alike. Waste collection, recycling and logistics could also benefit from tools such as product passports, logistics data and environmental footprint information to drive the circular economy. These tools would likewise be based on large amounts of cross-domain data. The changing climate is another key challenge where data can enable better decision-making. For instance, existing infrastructure is especially affected by changing weather patterns: flood risks, mobility, urban logistics, etc. need action not only in the short term to limit and control damage, but also continuous collection of data to drive policies in the longer term.

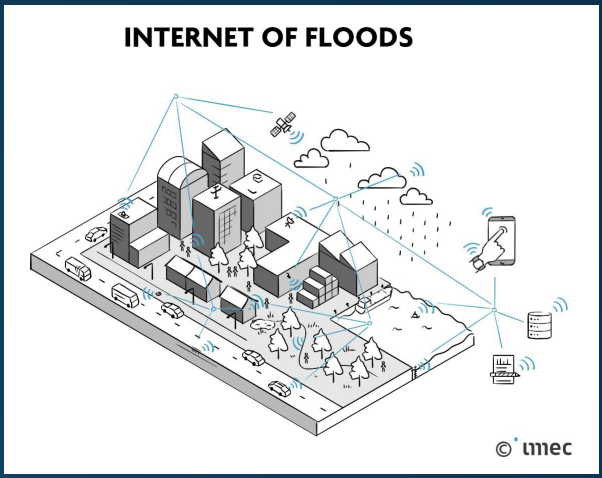
Figure 1: Examples of major societal challenges



The illustration in the box below gives a concrete example describing the 'Internet of Floods' (IoF) case study from the University of Ghent [6], where data from existing and new technologies can be highly valuable in addressing complex challenges.

IoF Case study: near real-time prediction of flooding.

The paper mentioned in "A Review of the Internet of Floods: Near Real-Time Detection of a Flood Event and Its Impact" [6] gives a summary of data that can be aggregated within an Internet of Floods (IoF) to assist the near real-time detection of a flood event and its impact. The paper describes the need to look for the optimal way to collect data in order to detect floods in real time, and it reports on the current state of research on IoT in the domain of flood detection. The integration of real-time IoT sources with other sources (also called data fusion) would greatly enhance disaster management. It concludes that the IoF could enable detection of a flood event and its impact in near real time, enabling the community to be immediately warned and informed. Different sources of data that could feed the IoF are shown in the figure below. These include pluvio meters, satellite images, weather monitoring stations, smart homes, cameras, buoys and sewer sensors, water level sensors in rivers and sewers, solicited and unsolicited crowdsourcing via e.g., social media channels, smart vehicle data (such as on-board data from windshield wipers and tyres), and much more. This example is a good illustration of how heterogeneous data from different origins unlocked by Society 5.0 could contribute to the early warning of flood risks.



Addressing societal challenges typically requires data from different domains to be considered. For example, optimizing a product's environmental footprint also calls for optimized urban logistics. In healthcare services, management of pandemics could be optimized via early discovery of inflammatory diseases among the population. In energy, optimized urban logistics could have relevance for enabling flexible energy demand and supply. And the early warning of flood risks described above could make an important contribution to overall urban disaster response. The list is vast, but all these potential applications have one challenge in common, namely the need for data to be shared at societal scale. A first good step towards that goal is to make cross-domain data Findable, Accessible, Interoperable and Reusable (also known as making data 'FAIR' for short), as set out in the following section.

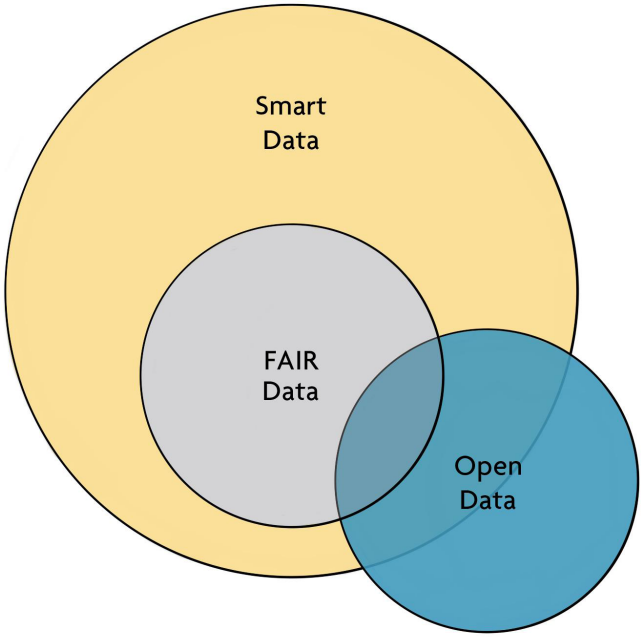
3. Making data Findable, Accessible, Interoperable and Reusable (FAIR) and ready to share

FAIR [5] data stands for Findable, Accessible, Interoperable and Reusable data. The term is widely used in the context of e.g., open science (research) and health data. GO FAIR [7] explains how to make data FAIR, defining the principles and sub-principles of each attribute. Making data FAIR focuses amongst other things on using unique and persistent identifiers, clear usage and definition of metadata, use of open protocols and community standards, use of well-established ontologies and vocabularies, and establishing key expressions on how to use the data.

The GO FAIR web pages state: *"The principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data."* Applying the FAIR principles is a first step in making data exchangeable and usable.

Making data FAIR is also a precondition for building a data catalog (as explained in *"Data Catalogs: A systematic Literature Review and Guidelines to Implementation"* [8]). In addition, catalogs need three other vital components namely: metadata management, business context and the operation of data responsibility roles such as data stewards. As described in the Data Catalogs paper, FAIR forms the vital foundation for publishing data and its metadata.

Figure 2: The relation between smart, FAIR and open data [5]



Source: https://www.ugent.be/en/research/datamanagement/after-research/fair-data.htm

While making data FAIR is a good first step towards sharing data at societal scale (the value of which was introduced in Section 2), FAIR data is not the final stage in making data ready for decision-making. The data also needs to be 'managed' to make it trustworthy for all stakeholders. Figure 2 (inspired by the University of Ghent presentation of FAIR data [5]) shows the relationship between FAIR, open, and smart data. The figure should be interpreted at the level of properties: FAIR data covers some of the properties of smart data, but not all of them. The rest of the paper focusses on the smart data attributes in the context of decision-making on a societal scale. The next section describes some of the key challenges in this type of decision-making.

4. Challenges in using data for decision-making

Data-informed or data-driven decision-making is the process of making decisions based on actual data rather than intuition, experience, or observation alone. Most professionals involved in decision-making know that bias and false assumptions (in the absence of trustworthy data) can lead to poor decision-making. There are many sources of bias in decision-making [9]. But even when using more 'objective' data, there is still the risk of it being subject to non-qualified or non-transparent data bias or quality issues (e.g., inaccurate, unreliable, ambiguous or erroneous data).

The simplest way of verifying the trustworthiness of data for decision-making is to measure compliance of the data to the outcome of the decision after it has been taken and use feedback loops to optimize the quality continuously. This can be relatively easy for simple operational decisions with low impact, but very complex for more strategic policy-driven and long-term decisions and might require contextual or continuous measurements. Using data in decision-making does not work with a big bang strategy, it needs to be approached step by step together with the stakeholders. That is the real meaning of smartness: the ability to apply what has been learned in incremental steps. Addressing data bias and quality is one of these important and continuous challenges. Let's use data bias as an example.

Data bias refers to the systematic distortion of data in a way that leads to inaccurate or unfair conclusions. It can occur in various ways, some of which are:

1. Sampling bias: This occurs when the sample of data used to make conclusions is not representative of the entire population. For example, if a survey on the political views of a country is conducted only among people living in urban areas, it may not accurately reflect the views of the entire population.
2. Selection bias: This occurs when the data used to make conclusions is not chosen randomly, leading to a distortion of the data. For example, if a study on the effectiveness of a new medication is conducted only on patients who are more likely to respond positively to the medication, it may not accurately reflect the effectiveness of the medication for the general population.
3. Measurement bias: This occurs when the data collected is not accurately measured or recorded, leading to a distortion of the data. For example, if a study on the height of a population is conducted using a measuring tape that is consistently one inch too short, it will result in an underestimation of the average height of the population.
4. Confirmation bias: This occurs when data is selected or interpreted in a way that confirms preexisting beliefs or hypotheses. This can lead to a distorted view of the data and can prevent the discovery of new or alternative explanations.
5. Algorithmic bias: This occurs when a machine learning algorithm is trained on data that is biased, leading the algorithm to make biased decisions or predictions. For example, if a facial recognition algorithm is trained on a data set that is predominantly composed of white faces, it may have difficulty accurately recognizing faces of other races.

In essence, bias is an inclination, prejudice or directionality to information, and it can creep into the data at any stage within the data pipeline, e.g., in the data capture, in semantic annotations, in transformations and algorithms, as well as in fusion or sharing.

Data bias introduced by machine learning algorithms.

Typically, machine learning algorithms themselves are very susceptible to data bias. The output from machine learning algorithms is rarely made explicit i.e., the details of what data was used, how it was generated and what was done to the data before modelling. This is a key issue in embedding these tools in data pipelines that feed decision-making processes, nicely captured in the visual below.

As illustrated above, a typical place where data bias can occur is in using machine learning algorithms aiming to increase the value of (combined) raw source data into higher value data. Some other typical examples of data bias in machine learning are algorithmic, sample, prejudice, measurement and exclusion bias [10].

A second example of data bias can be the presence – and unmanaged sequenced transformations of – semantics in a data pipeline. When data traverses from its origin through data processors to a delivery point in a marketplace, the meaning of the data can be modified or redefined along the way. Data can have different semantics ranging from business and technical semantics to regulatory ones. Unmanaged semantic mappings (which may also be combined with erroneous or inaccurate mappings) can easily introduce bias into the system due to different interpretations.

IoF Case study reflection: bias introduced by unmanaged semantics.

The case study introduced above shows that a wide range of data sources can contribute to the creation of the Internet of Floods (IoF) and assist in near real-time flood prediction. Each of the data sources produces different kinds of data, from smart sewer systems, water level sensors, smart vehicle data, weather monitoring, remote satellite sensing systems to unsolicited crowdsourcing on social media. The meaning of the output data will be very different and sensor-specific, but it will all contribute to a single, unique IoF and to an accurate prediction of how much water will be at a certain place at a certain time. In generating the IoF, data will thus need to be transformed, mapped, refined, etc. This also means that bias can creep into these processes, and into interpreting the meaning of the input data and defining that of the output data for the next stage. Managing the semantics and semantic mappings is key to being able to assess the risk of bias and to allowing final stakeholders to continuously review the data pipeline. It is important to understand that at one moment a data point could be assigned a particular meaning, then subsequently it could be redefined, or even equated to the same thing from a different source.

A third example of bias can be described as response bias. This is common on social media platforms where a small number of active users produce the largest number of posts.

IoF Case study reflection: response bias using unsolicited social media.

The case study shows that unsolicited media, such as live-time images or text about floods from e.g., Twitter, can make a real contribution to near real-time predictions of floods. However, not all regions of a city have the same digital maturity or access to social media platforms or typical users of these media. Some neighborhoods can be very active, while others may have far fewer or no active users. How do we make sure that this type of data in the IoF does not cause bias in the eventual decisions taken (e.g., in assigning priorities for interventions of the emergency services) and still deliver fair services to all citizens?

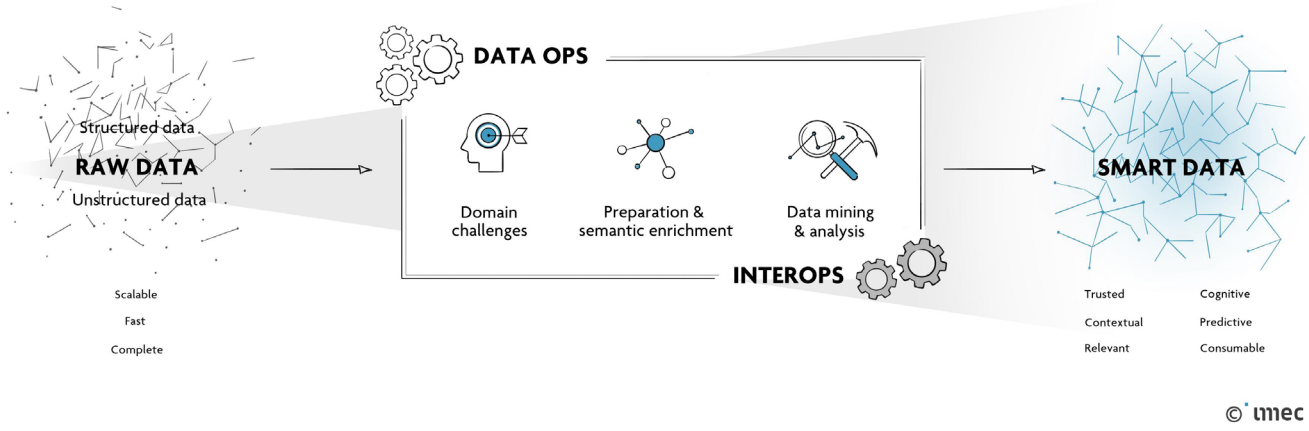
This section introduced some examples of data quality issues that can threaten trust in the quality and usability of data for decision-making. It focuses on data bias as this is an important element to consider when using new technologies that enable the collection and analysis of combined big data sets. There is clearly a need to have detailed insight into how the data were generated and processed, and how they were used in the process leading to the final data set which is provided to the decision makers. The next section discusses how smart data principles are at the heart of controlling the necessary quality of data.

5. Managing data at societal scale using smart data principles

5.1. Conceptual introduction to smart transformations of data

Given that data for data-informed decision-making at societal scale calls for minimal data bias, and maximum understanding and transparency, how can these goals be achieved? The answer is hinted at in Section 2: the solution lies in using managed data at scale i.e., smart data, data which is produced at scale and that is ready for decision-making. Smart data is the result of transforming raw data by applying the FAIR principles, making it fit for purpose (e.g., addressing bias in decision-making, transforming it into efficient processable formats, ...), and governing the processes used along the way as illustrated in Figure 3.

Figure 3: Smart data focuses on essential process steps, which can be executed within interoperable and standardized data operations



© imec

The first step within the smart data generation process is to understand the domain challenges (which can be societal or business) and the associated use cases. The real goal for smart data in the context of better decision-making is to enable actionable insights that can be used in a specific (or multiple) domain(s). A thorough knowledge of the reasons for using the data (e.g., identifying applications which will use the data) within a particular domain is the basis for determining the correct next steps in creating smart data from large quantities of raw data. Expressing these domain challenges in a standardized way is key to the societal scale mapping of data to a broad usage. Section 6 describes how the VLOCA [11] and OSLO [12] (Open Standards for Linking Organizations) programs in Flanders are standardizing this.

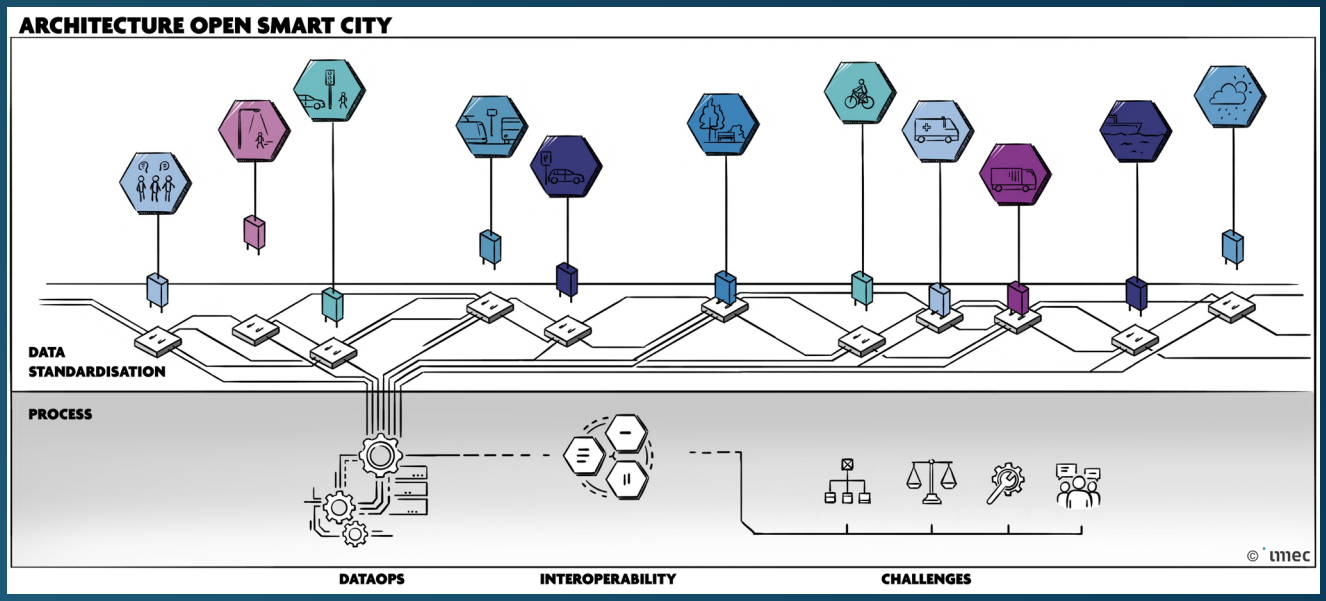
A checklist of smart data properties can be a useful tool when addressing domain challenges in decision-making (whether on an operational or policy level). Such a checklist, introduced by James Kobelius [13], is described in Section 5.2 below. This list classifies properties to target as Trusted, Contextual, Relevant, Cognitive, Predictive and Consumable

Having identified the domain, the next steps are data preparation, semantic enrichment, data mining and analysis. These steps can be sequenced multiple times and are essential to converting sets of raw data into well-prepared, well-understood data sets and streams. The process of managing and governing the different steps is driven by what are known as 'DataOps'. To be able to do this in a reproducible and scalable way across organizations, requires 'InterOps' (defined as using standards and best practices as much as possible).

Definition of DataOps.

We introduced DataOps in the white paper on Open Smart City Architectures in 2019 [14] along with the Gartner definition of DataOps [15]: **"DataOps is a collaborative data management practice focused on improving the communication, integration, and automation of data flows between data managers and consumers across an organization. Much like DevOps, DataOps is not a rigid dogma, but a principles-based practice influencing how data can be provided and updated to meet the need of the organization's data consumers. The goal of DataOps is to create predictable delivery and change management of data, data models and related artifacts. It uses technology to automate data delivery with the appropriate levels of security, quality, and metadata to improve the use and value of data in a dynamic environment."**

DataOps was introduced at that time as one of the basic building blocks for creating a smart city, as illustrated in the figure below.



Data preparation and semantic enrichment refer to the set of processes used to clean and transform the raw data prior to processing and analysis in the following stage. They involve detecting errors and making corrections to the data (e.g., removing outliers), combining data sets if needed, reformatting data, assigning (and standardizing) data formats, enriching source data, assigning a unified meaning (semantics) to the data, etc. These steps are not only needed for qualitative processing; they are also vital elements of data management that enable data governance. Especially when preparing data to be shared within an ecosystem of data pipeline actors, these steps are mandatory. The preparation and semantic enrichment need to be aligned to the domain challenges, and they must be conducted with consideration to their position in the data chain. As discussed in Section 4, bias in semantics is a significant risk. It is thus important to track semantic mappings and enrichment e.g., using knowledge graphs and information models and to offer these as metadata to the data pipeline clients.

Data mining and analysis refer to the set of processes that operate on prepared data sets and that search for and create patterns, trends, connections, aggregations, links, etc. The goal is to extract more value from the data, i.e., useful information that can be shared with the applications. This DataOps and InterOps focus also allows for data governance, so that the processed data can still be linked back to the raw data and its original meaning. This is important, for instance, when machine learning is used to detect clustering of data. The original raw data and its quality may change continuously, and new raw data sets may be added on the fly. Thus, while the flow of information may seem steady, its quality may change continuously.

5.2 Target data attributes when applying smart data operations

The activities and processes described above target the creation of smart data (managed data ready for decision-making at societal scale), and they can be evaluated against the checklist of six main properties below [13].

ATTRIBUTE	DESCRIPTION
Trusted	To be able to take decisions on insights obtained from data, having trust in the quality, consistency, accuracy, origin, protection and security, etc. of that data is crucial. A defined level of trust is needed to enable accountability for the actions that are taken based on the data.
Contextual	The context of data is an important part of the metadata of a data collection. It gives information about the collection of the data (e.g., location, time, measurement conditions, etc.); the use of the data (usage policies, key interpretation guidelines, constraints, limitations, etc.); the managed meaning of the data and much more. The context of the data should be sufficiently complete to address, as far as possible, the full range of use cases. The more domains that the data can address, the more context will be needed. Notice that different contexts of data could point to each other, e.g., using semantic web linked data techniques. Data provenance (i.e., records about the trail of the data) is also part of the context and is key to enabling trust.
Relevant	Even when data are trusted and contextual, they are not necessarily directly usable or relevant for the targeted use cases. It is important to organize the data from different sources in such a way that applications can easily find them, and that the data are published and presented in such a way that they are directly relevant to the applications. For that, data needs to be curated.
Cognitive	In many cases, raw data needs to be processed to make it more understandable, especially when it is unstructured e.g., streaming media, images, etc. Cognitive computing refers to technology such as machine learning, machine reasoning, natural language processing, speech recognition, and human-computer interaction that transforms data into more usable forms for economic or decision-making purposes.
Predictive	In the process of transforming raw data into data that provides insights to drive certain decisions, it is important to focus on trends and predictions. For example, in digital twins it is important to be able to make predictions about the evolution of certain data sets such as flood risks or the impact of certain actions, and to go back and forth in time. Smart data should be tailored and presented in such a way that prediction is facilitated, e.g., by adding enough context of certain causes of trend changes in the data and allowing for 'time travel'.
Consumable	Data needs to be presented so that it can be easily used by humans and machines. Thus, it is very important to design the access, presentation and service layers (e.g., API) so they are appropriately tuned to the usage that will be made of the data. To make data consumable, the needs of the decision makers and their tools (e.g., digital twins) need to be well understood and targeted.

5.3 Summary of important smart data principles and minimum mechanisms to pursue

To deliver smart data at societal scale, there are some highly important (but not necessarily exhaustive) principles as shown in the following table.

PRINCIPLE	DESCRIPTION
Make Fit-For-Purpose	Use a structured process to detect and describe the domain challenge(s), not only from the functional perspective but also from an understanding of the risks of using data to address the challenge. Translate this understanding into a description of the targeted smart data attributes, which should be decoupled from the final application(s). This allows data to be fit-for-purpose from the beginning.
Manage Data and Algorithm Semantics	Data semantics and their mappings need to be exposed (e.g., using knowledge graphs), as they are key to understanding how the data was transformed and interpreted along the way. Semantics not only play a key role in describing the data, but also in describing the metadata of the algorithms that create data.
Standardize DataOps	Understand how the data was generated and analyzed along the way, as this is the basis for addressing data bias and quality. Standardized DataOps processes and artefacts across the ecosystem are needed to do this at scale, i.e., describe how the data was prepared, enriched and mined. In understanding how the data was generated, a clear separation of data and algorithm management (made explicit in asset management) is needed. Standardization is crucial to be able to share this understanding with the data users.
Protect data	Ensure data protection by allowing data controllers to act independently from other data actors (e.g., by supplying private data vaults or intermediate agents) and provide mechanisms to 'move the algorithms, not the data', i.e., run algorithms close to where the data are stored, for instance on the same server(s), rather than transferring large amounts of data to an algorithm running on a remote device.
Decentralize	Targeting smart data for societal decision-making at scale needs a decentralized infrastructure as this opens standardized and controlled access to several sources of data and algorithms in order to, amongst other things, quickly validate data quality and applicability, and make comparisons.

These principles are highly analogous to the principles described in the Data Centric Manifesto [16]. The next section illustrates some of these principles in action.

6. Introduction to major Flemish and international initiatives that apply smart data principles

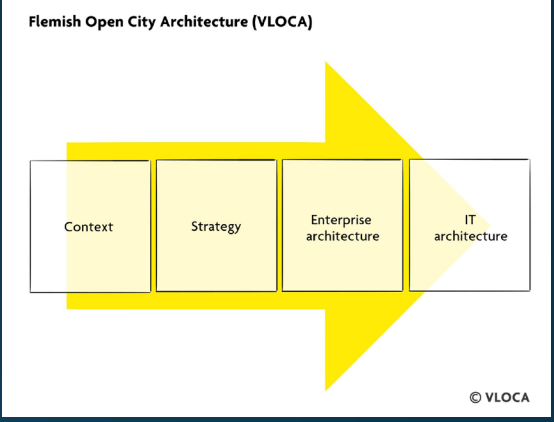
In this section, we illustrate how the principles introduced in 5.3 can be, and are being, put into practice, with examples from current initiatives in the Flemish and European landscape.

6.1 Make fit-for-purpose

An important first step in applying smart data principles is to understand the domain challenges and translate them into application-decoupled requirements for the smart data attributes. Common practices that enable the value of data to be unlocked for use within societal contexts are crucial. As an example, in Flanders, the VLOCA [11] program assists in capturing needs in a uniform process and is offered as a service to smart cities and communities. For its part, OSLO [12] (see 6.2) brings together a representative set of domain stakeholders to align on the definition of semantics and how to use vocabularies, technical standards and profiles for data exchange.

The VLOCA program assists in identifying domain challenges and needs in a uniform way

The VLOCA program aids cities and communities in tackling smart city challenges. One of the deliverables is a standardized approach in addressing and translating the needs of smart city projects, as explained in [17]. This is a very good way to understand and translate the domain challenges described in chapter 5 as base for a smart data approach, within its context, strategy, enterprise and data/IT architecture.



To address the needs of concrete smart data use cases such as these, imec has developed a smart data use case canvas (under continuous development). See Figure 4. This living document seeks to map the most important elements and identify the most critical assumptions to advance the innovation management process [18]. The canvas supports its users in reflecting critically on the stakeholders involved, their needs and the required data, along with the key resources to enable a given use case.

Figure 4: The imec smart data use case canvas (latest version at time of publication)

Stakeholder			
Needs / Opportunities			
Current Practices			
Currents Datasets / Models			
Jobs-to-be-done			
Value Creation			
Key Resources			
Trustworthiness Requirements			
Barriers			

imec

6.2 Manage semantics (and their related elements)

Unifying the meaning of data within important societal domains is key to being able to make data smart. Within the Flemish OSLO program [12], the combination of data vocabularies and ontologies combined with linked data principles and technologies is done at regional scale. The semantics are inspired by European standards but translated/supplemented within a regional context. Explicitly linking semantic transformations to the OSLO vocabulary and application profiles is a good way to start managing semantics. This can enable the ecosystem to converge and allow the data to be used, for instance, within a specific data space or data marketplace.

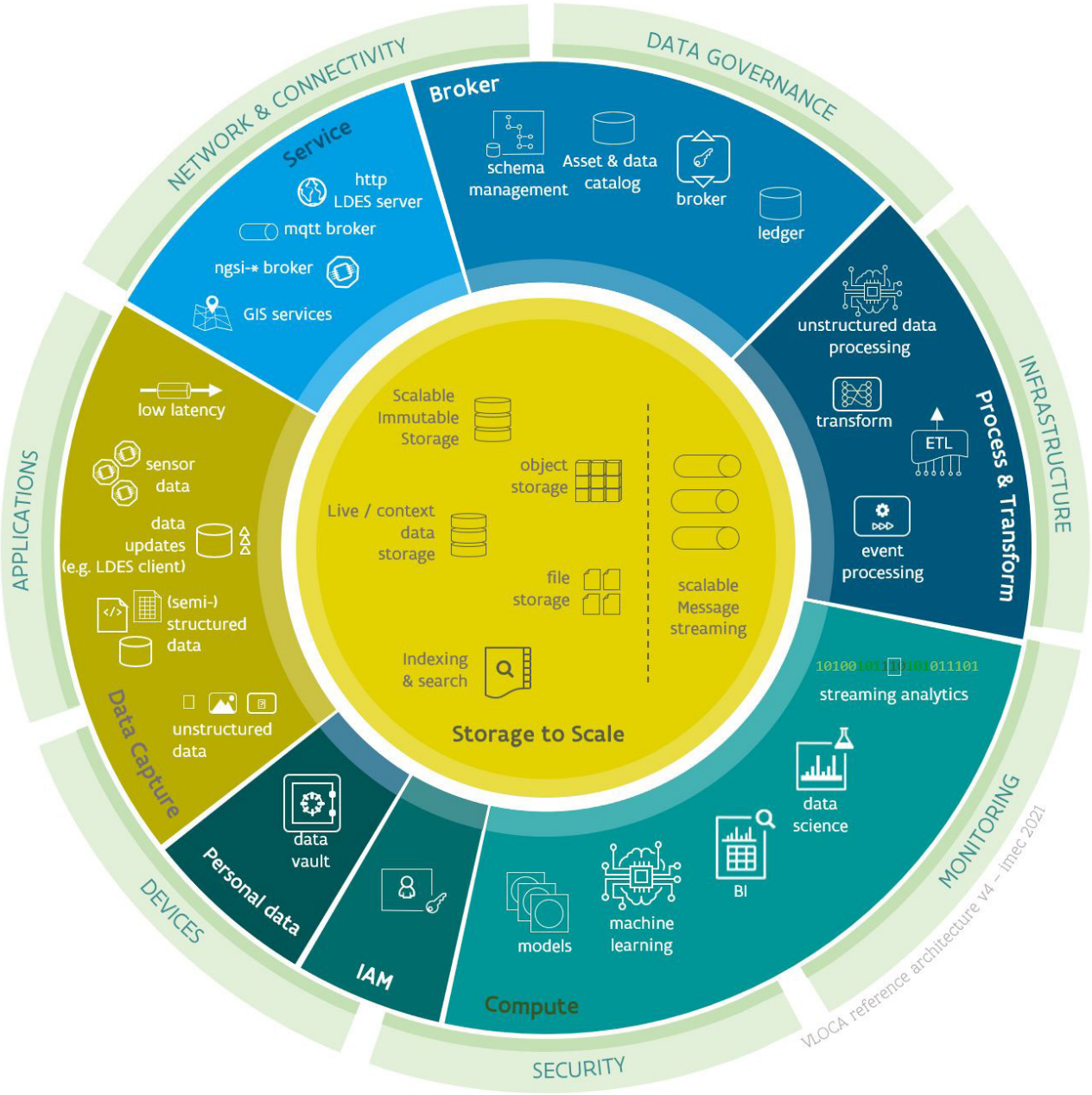
Semantics is not only important for data alone. It is also of crucial importance for managing compute and algorithm metadata such as standardized descriptions for machine learning algorithms that list the input and output smart data sources. Moreover, the characteristics of the alogrithm operation itself are becoming key targets for the smart data operations.

This need for transparency in using algorithms within decision-making has also been explicitly listed as Minimum Interoperability Mechanism (MIM) 5 by OASC (Open & Agile Smart Cities) [19]. This MIM addresses procedural and technical transparency, technical 'explainability', fairness, and context. These could be translated into a metadata API that every vendor would provide when supplying algorithms that can lead to high-impact decisions to cities, and that buyers could put in their procurement requirements. Ontologies and semantics for algorithm registers are currently in the process of being defined and standardized amongst cities, as can be found on GitHub [20].

6.3 Standardize DataOps

Data platforms exist in different flavors. The VLOCA program has defined a reference architecture [21] for a data platform and its distributed/ federated format. The goal is to facilitate discussion of generic DataOps technical processes and tools amongst different partners in the field of data management and operations. This 'VLOCA Donut' is shown in Figure 5 below.

Figure 5: The VLOCA Donut representing common technical capabilities of a smart data platform



The donut shown above can be seen as a group of technical capabilities to assist the transformation of raw data into smart data, and to share these with applications. The scope and definition of the different segments can be consulted on the VLOCA knowledge hub [17]. These technical functionalities assist in describing standardized DataOps processes: converting data for domain challenges, preparing and curating data, and allowing data analysis and mining to be integrated. Some (non-exhaustive) examples of functions that address the various attributes are given below.

Trusted:

- [IAM] Identity and Access Management allows authentication and can authorize access to specific resources.
- [Personal Data] Personal Data vaults allow individuals to take the role of data controller giving consent to authenticated users to access privacy or business sensitive resources.
- [Broker] Schema management, brokers and data catalogs can offer insights into the evolution of data resources and their metadata (e.g., context and meaning). Distributed ledgers can bring decentralized trust to transactions and contracts. Versioned schemas of data and compute resources (e.g., the model signature) allow time travel, which gives users transparent insights into the variability of the resources. Asset registries bring together a.o. ownership information, data schemas, information on transformation agents and data mining functions, and allow data lineage and provenance to be monitored.

Contextual:

- [Data Capture] On capturing data, the original raw data and its source context can be stored in the data lake as a reference to the original context.
- [Broker] Data Schema management allows for context to be added incrementally to the original raw data schema and version changes.
- [Process and Transform] Extract, Transform and Load functions allow addition of context to the data, or they make sure that the context of data transformations (e.g., data fusion) finds its way into the data scheme management and asset registries.

Cognitive:

- [Compute] To make data easy to understand, the use of compute elements such as streaming analytics, machine learning, etc. are essential parts of the smart data pipeline. They need to be fully integrated into the DataOps process, and also to be exposed with their own semantics and schemas so that they can be brokered.

Predictive:

- [Storage to scale] To be able to assess the behavior of data in time, time travel not only of the data values themselves is important, but equally of the synced changes in the data types and semantics. Time series databases, message streaming, and event sourcing are essential techniques for the creation of historic insights that allow trend predictions.

Relevant:

- [Storage to scale] and [Process and Transform] allow data to be formatted and indexed to service needs.
- [Service] Context is not only an internal object; it needs to be exposed to the client. Context exposure, management and brokerage are key properties of a service layer.

Consumable:

- [Service] The data can be offered in various ways – in time series, as immutable updates (e.g., linked data events), or linked with their context using international standards, protocols and tools such as MQTT, REST APIs, GRAPHQL queries, NGSI context APIs, LDES streams, ...

An essential element that is implicit within this donut is the need for data governance, and the ability to orchestrate components in a location-agnostic way (fog/cloud/edge) controlled by different organizations. Achieving this requires standards that ensure the smart data properties can be delivered over data pipelines which may be supplied by multiple different vendors.

6.4 Protect data

Data protection entails an explicit separation between services and control of the data they use. This is necessary for the protection of privacy in personal data and of intellectual property in commercial data. The SOLID program [22] in Flanders is an example of enabling personal data protection by giving data subjects direct control over the use of their data for value-added services, with a clear emphasis on interoperability and data portability. The program provides personal data vaults to keep data separate from these services and a real-time consent-based access control mechanism under the control of the data subject. These linked-data based vaults and the managed personal data space ecosystem around them are provided by the Datanutsbedrijf (the Flemish Data Utility Company) [23].

Data protection is also covered by the OASC Minimum Interoperability Mechanisms. In particular, MIM4 deals with trust [24], addressing data protection from two perspectives: "That of Individual citizens in terms of transparency & privacy preferences collection", and "Cities and Data Using Services (Data Controller/Processors/) in terms of Authorization and Data usage control and enforcement".

6.5 Decentralize

Decentralization is at the heart of datafication at societal scale. Use cases and applications need fast, easy, trusted and sovereign access to multi-organization data and compute assets to help them innovate, and give them more means to address data insight and quality issues. The rollout of data spaces has become a strategic goal for the European Union and will boost the decentralized nature of data sharing at societal scale.

Some key programs that illustrate this are:

- The VSDS (Vlaamse Smart Data Space) [25], which is part of the Flemish 'Relance plan'. It uses European standards to build software components that support decentralization within data spaces. As described in the illustration below, data spaces are at the heart of the European Union's strategy.
- The VLOCA program. In co-creation with cities and communities, VLOCA has defined a virtual smart data platform within its Open Architecture [21].
- The European research program DUET (Digital Urban European Twins) [26] has defined a decentralized T-Cell architecture for digital twin components such as visualization, simulation models, IoT data platforms, etc. to leverage the deployment of cross-domain digital twins.

Introducing European data spaces.

International organizations such as Gaia-X [27] and IDSA [28] represent industrial needs and they have driven the definition and adoption of data spaces over recent years. In February 2020, the European Commission published its data strategy [29] expressing the need for a single market for data, where data spaces play an important role.

The OpenDei Design principles define data spaces as [30]: "A data ecosystem, defined by a sector or application, whereby decentralized infrastructure enables trustworthy data sharing capabilities."

And very recently, at the beginning of 2023, the DSSC (Data Spaces Support Centre) [31] issued its Starter's Kit to assist the ecosystem.

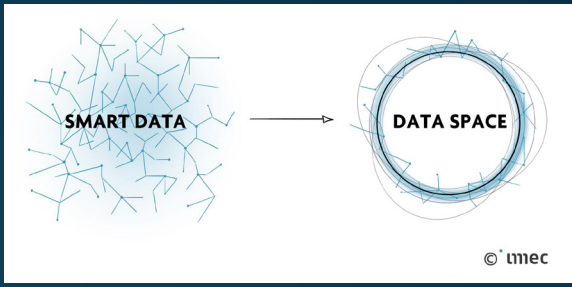
Why data spaces need smart data.

The OpenDei [30] document mentioned above also includes a more detailed definition of a data space, namely:

"From a technical perspective, a data space can be seen as a data integration concept which does not require common database schemas and physical data integration but is rather based on distributed data stores and integration on an "as needed" basis on a semantic level. Abstracted from this technical definition, a data space can be defined as a federated data ecosystem within a certain application domain and based on shared policies and rules. The users of such data spaces are enabled to access data in a secure, transparent, trusted, easy and unified fashion. These access and usage rights can only be granted by those persons or organizations who are entitled to dispose of the data."

"Realizing interoperable data spaces is more of a coordination challenge: agree on standards and design principles that are accepted by all participants. While making data interoperability work in pilot applications, proof of concepts, and living labs is relatively easy, the real challenge lies in viewing interoperability as the new norm for facilitating mass adoption and scalability."

Data spaces facilitate data sharing in business ecosystems close to the domain applications. For that, data needs to be prepared for sharing and use, and be ready to apply standard practices on legal, technical, and operational levels. Preparing data to be smart facilitates this as it is ready for scalable business usage (and thus has economic value), it is prepared to be trusted (one of the key aspects of data spaces), and its metadata and context have been consolidated from a usage level, so that e.g., standard ontologies and vocabularies can be applied to them.

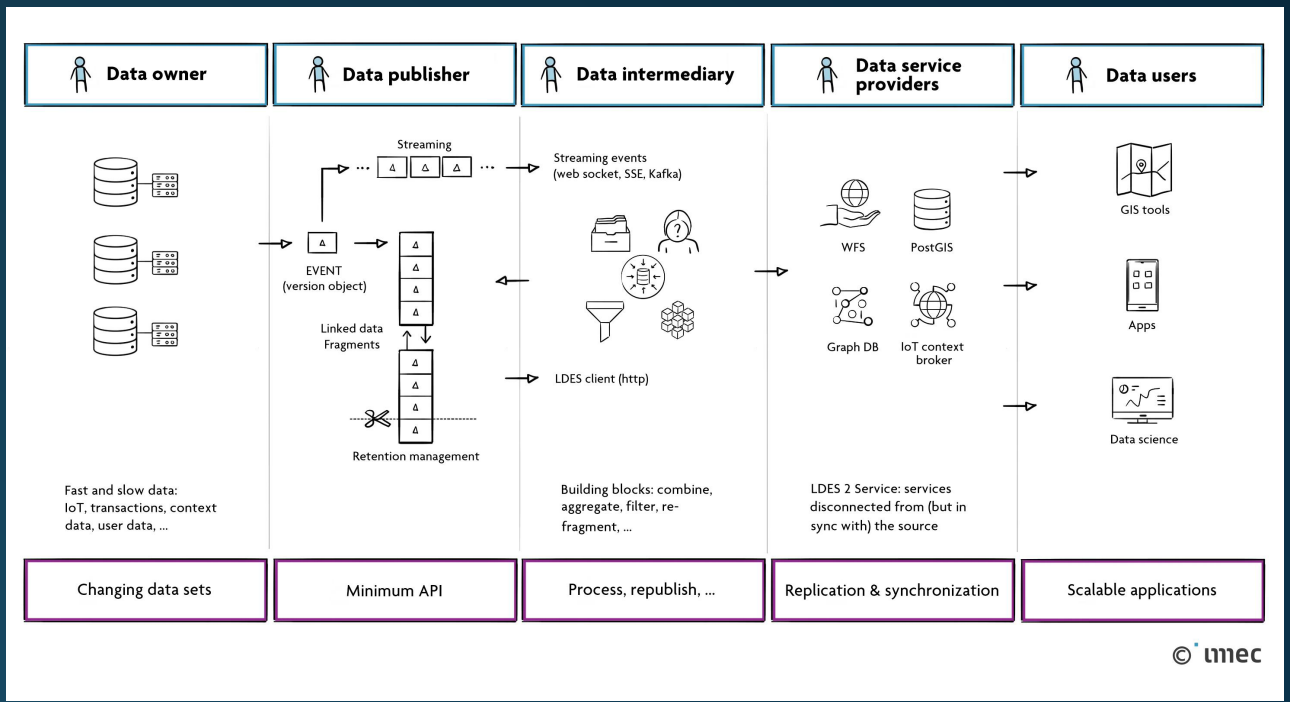


Vlaamse Smart Data Space (VSDS) [25].

The OpenDei document mentioned above also introduces the concept of a 'soft infrastructure', which is defined as an invisible layer made up of technology neutral agreements and standards on how to participate in an ecosystem. Examples of soft infrastructures are the global GSM network, payment networks and the Internet. They are all based on agreements to be implemented on 'hard' infrastructures and adopted by users in a decentralized and competitive fashion.

The soft infrastructure provides a level playing field for data sharing and exchange and specifies common functional, legal, operational and technical aspects such as security, identity, authentication, protocols, metadata, etc. In order to establish sector-adequate data spaces, the soft infrastructure needs to be complemented with sector-specific aspects.

The VSDS program [25] aims to deliver important standards and tools to support the creation of soft infrastructure. Recently, Digitaal Vlaanderen celebrated the 10th anniversary of OSLO [12] which delivers metadata standards (vocabularies and ontologies) and application profiles to assist the semantic standardization process. The VSDS program complements this with standards for publishing, finding and consuming data (e.g., the Linked Data Event Streams (LDES) [32] protocol) in a scalable and secure way using linked data principles and balancing costs at the production and consumption sides. These two programs deliver standard ways to handle cross-domain interoperability (data models and exchange, data exchanges, APIs and protocols, provenance and traceability) that can easily be implemented on existing infrastructure such as existing data platforms that want to participate in data spaces. The figure below illustrates how LDES enables scalable data publishing. This way of publishing data allows several key roles (data owner, publisher, intermediary, service provider and user) to work efficiently on the data, maximize value, minimize overheads and leverage a scalable landscape. This maps perfectly onto the ambitions of data spaces. To account for the provenance of smart data streams (e.g., using LDES), the Smart Data Specification indicates how this could be standardized [33].



7. Conclusions

As we identified at the start of this paper, the term 'smart data' is commonly used in the context of decision-making. However, its definition does not refer so much to a particular type of data, rather it refers more to the processes and techniques that can be used to transform data step by step towards a state where they can be used better in the process of data-informed decision-making. If data and their processing are not understood and not well-described, and they still need a lot of processing at the application side, they cannot be called smart. Making data fit-for-purpose introduces data transformations using e.g., algorithms that are essential parts of the smart data pipeline. It also needs to be made transparent within its context to induce trust at the consumer level.

Data is already being used extensively in decision-making. But it does not happen at scale, especially in public environments where data come from very heterogeneous sources and the problems are cross-domain. However, the power and widespread adoption of recent and emerging technologies such as AI, IoT, blockchain, big data, etc. to capture, store, process and use data – along with the rising complexity of public-domain challenges – give rise to Society 5.0 (see Figure 3), defined as "A Human-centered society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space". In this context, digital twins for public government are rapidly gaining momentum as tools to integrate cyberspace and physical space and to assist in decision-making at scale. However, they will need smart data to do this.

In this paper, we tried to sketch what smart data is or could be by pointing to its main attributes and processes, and the methods to realize them. We also gave some basic principles to adhere to for obtaining reliable decision-making, which is clearly not solely a problem of the decision maker but needs a holistic approach and cooperation between the different data pipelines and data exchange points. Moreover, it is not only the pure data that needs to be made smart, the algorithms that transform data also need to apply the smart data principles, as more algorithms will move towards the data in a decentralized and unified way, rather than the traditional flow of data towards the algorithms.

Using data at scale is not only a technical challenge. It is also a cultural one, where new roles and responsibilities are needed to increase control and accountability. These roles can be data managers and engineers to setup the Data Ops workplace and shared dashboards, data stewards to apply data governance, business professionals to assist in data search and collaboration, and system actors (e.g., companies delivering information systems and applications) to offer rich open APIs for convenient usage. We need to work smart(er) with data from the start and face the challenges together with all the actors who play a role in the scalable and durable resolution of the problem.

We also tried to illustrate that a lot is already moving in the digital and data field within Europe and Flanders to assist this. The smart data train(s) are moving! On standardization, many promising results can also be expected (in addition to the plethora of existing standards e.g., communication protocols, data formats, big data stacks, event buses, ...). These will be on:

- algorithm metadata standardization [20] driven by the cities of Amsterdam, Helsinki, Barcelona, etc. and OASC within the context of algorithm registers
- dataspace standardization by Gaia-X, IDSA, DSSC (not only on technical level but also on business, legal, governance levels) and also in Flanders with VSDS
- personal data protection using e.g., SOLID (Flanders)
- domain data standards by OSLO (Flanders), Fiware (smart data models), W3C, OGC, ...
- new MIMs being defined by OASC and mapped to existing or new implementations
- and many more.

So, the time is right to look at the various processes and examples and determine what all this could mean for your organization, whether you are a data or algorithm provider, a decision maker, a data processor, or other. New technologies like AI are leading to rapid innovation. Just take the example of ChatGPT [34], an intelligent and natural chatbot (based on large quantities of data fed into a deep learning algorithm) that can answer your questions. But how can we trust the answers? In Google Search, we still have verifiable links to determine for ourselves if we trust the results, with ChatGPT this is currently not possible. Such developments underscore that the need for smart data will be greater than ever in the coming years, especially in the context of making game-changing decisions in a societal context.

References

[1] **Datafication:** *"Datafication is a technological trend turning many aspects of our life into data which is subsequently transferred into information realized as a new form of value"* from https://en.wikipedia.org/wiki/Datafication

[2] **Local Digital Twins:** https://vloca-kennishub.vlaanderen.be/Open_urban_digital_twins

[3] **Society 5.0:** https://www8.cao.go.jp/cstp/english/society5_0/index.html

[4] **Smart data:** Baldassarre, M. T., Caballero, I., Caivano, D., Rivas Garcia, B., & Piattini, M. (2018, November). From big data to smart data: a data quality perspective. In *Proceedings of the 1st ACM SIGSOFT International Workshop on Ensemble-Based Software Engineering* (pp. 19-24).

[5] **FAIR data (1):** https://www.ugent.be/en/research/datamanagement/after-research/fair-data.htm

[6] **A Review of the Internet of Floods: Near Real-Time Detection of a Flood Event and Its Impact:** Van Ackere, S., Verbeurgt, J., De Sloover, L., Gautama, S., De Wulf, A., De Maeyer, P. (2019, October) https://www.mdpi.com/2073-4441/11/11/2275

[7] **FAIR data (2):** https://www.go-fair.org/fair-principles/

[8] **Data Catalogs:** A systematic Literature Review and Guidelines to Implementation https://www.researchgate.net/publication/354697372_Data_Catalogs_A_Systematic_Literature_Review_and_Guidelines_to_Implementation

[9] **Bias in Decision Making:** https://timelyapp.com/blog/decision-making-biases

[10] **Bias in Machine Learing:** https://towardsdatascience.com/5-types-of-machine-learning-bias-every-data-science-should-know-efab28041d3f

[11] **VLOCA:** https://vloca.vlaanderen.be/

[12] **OSLO:** https://www.vlaanderen.be/digitaal-vlaanderen/onze-oplossingen/oslo

[13] **Smart Data:** https://www.dataversity.net/big-data-smart-data-big-drivers-smart-decision-making/

[14] **Open Smart City Architecture:** https://www.imeccityofthings.be/drupal/sites/default/files/inline-files/open_city_vision_paper_final.pdf

[15] **DataOps:** https://www.gartner.com/en/information-technology/glossary/dataops

[16] **Data-Centric Manifesto:** http://www.datacentricmanifesto.org/

[17] **VLOCA Knowledge Hub:** https://vloca-kennishub.vlaanderen.be/VLOCA_traject

[18] **Innovation Management Process** Schuurman, D.; Wuyts, G. & De Meester, T. (2022). Living Labs for scoping Digital Twins: introducing imec's Innovation Management approach. In Proceedings of the Open Living Lab Days 2022.

[19] **OASC MIM5 on transparency:** https://mims.oascities.org/mims/oasc-mim5-transparency

[20] **Standardization of algorithm registers:** https://github.com/Algoritmeregister/standard

[21] **VLOCA reference architecture:** https://vloca-kennishub.vlaanderen.be/Open_Smart_City_Architectuur

[22] **SolidLab Vlaanderen:** https://solidlab.be/

[23] **Datanutsbedrijf:** https://www.vlaanderen.be/digitaal-vlaanderen/het-vlaams-datanutsbedrijf

[24] **OASC MIM4 on trust:** https://mims.oascities.org/mims/oasc-mim4-trust

[25] **Vlaamse Smart Data Space:** https://www.vlaanderen.be/digitaal-vlaanderen/onze-oplossingen/vlaamse-smart-data-space

[26] **DUET(Digital Urban European Twins):** https://www.digitalurbantwins.com/technical-approach

[27] **Gaia-X:** https://gaia-x.eu/

[28] **International Data Spaces Association:** https://internationaldataspaces.org/

[29] **European Data Strategy:** https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

[30] **Design Principles for data spaces:** https://design-principles-for-data-spaces.org/

[31] **DSSC (Data Spaces Support Centre):** https://dssc.eu/

[32] **Linked Data Event Streams:** https://semiceu.github.io/LinkedDataEventStreams/

[33] **The Smart Data Specification:** https://biblio.ugent.be/publication/8771000/file/8771003.pdf

[34] **ChatGPT:** https://openai.com/blog/chatgpt/