Data Spaces in Denmark

A vision by the Nordic IoT Centre

Nordic IoT Centre

FOUNDED BY ALEXANDRA INSTITUTE AND FORCE TECHNOLOGY

***Alexandre Alapetite**, Principal Software Solutions Architect,*
AI and Data Analytics, Alexandra Institute (alexandre.alapetite@alexandra.dk)

***Rina Vijayasundaram**, Research Assistant,*
Aarhus University (rina@cc.au.dk)

***Martin Brynskov**, Lector,*
Aarhus University (brynskov@cavi.au.dk)

***Anders Struwe Mynster**, Head of Department*
IoT, Data & Services Innovation, FORCE Technology (apm@force.dk)

***Mads Johansen**, Specialist,*
IoT, Data & Services Innovation, FORCE Technology (majh@force.dk)

Financed by:

WE BUILD DENMARK

1 Introduction

Data is becoming a huge part of society, trends such as Big Data, AI, IoT, Industrial IoT, and so on, are becoming more and more integrated with society. But how does one go about using data for new applications or services? When an individual, a start-up company, a medium-sized company, or even a large enterprise, is starting a new project that requires data and they do not have it themselves or the resources to collect it, where do they go to it then? There is much data available from different organizations and companies – both open data and data available for purchase. But even though the data is available somewhere, it is not always easy to find what is needed for a specific project. And when the data is finally found, how does one make sure, that the planned use of that data is within the terms of services and regulations?

This white paper gives a description of data spaces. The current state of data spaces alike a vision of the future – through trends in the research and political initiatives. The white paper aims at giving individuals, as well as companies, interested in using open data, a brief overview of where to find data sets in the current landscape. It also describes a few initiatives set in motion by different organizations as well as how the landscape looks regarding key actors and policies. The focus is on Denmark's data spaces and actors in this area. Furthermore, there will be a focus on the Smart City aspect of data spaces. Examples will be given for current data spaces and data catalogues in Denmark, as well as examples from around the world.

Regarding the terminology, this white paper sees the continuity between "data catalogues", "data portal", and "data spaces" and understands "data catalogues" as a simple list of datasets. "Data portal" is a more evolved list of datasets including indexing, searching, potentially data previews. Finally "data spaces" is the most current evolution, including both more potent technical tools as well as softer aspects such as governance and stricter rules.

2 Definition of Data spaces

There is no stable definition of data spaces yet, but a simple simplified answer could be: "*Data Spaces are collections of data sources, including definitions of the data, shared between multiple parties, typically publicly available, and with a description of the terms of use including possible financial payments.*"

The International Data Spaces Association (IDSA) has a more detailed definition, which adds some specific terms, which will be described more in-depth.[1]

> "*A data space is defined as a decentralized infrastructure for trustworthy data sharing and exchange in data ecosystems based on commonly agreed principles.*"
>
> IDSA, Position paper, "Design principles for data spaces", April 2021

An argument for the data spaces to be **decentralized** is to keep companies and individuals who generate data in control. Thus, the data is not controlled by the data space, it is simply shared and exchanged through it. This is closely linked to data sovereignty which is central for such data spaces and will exactly describe the owner's control of the data. It is defined as: "*a natural person's or*

[1] OPENDEI, "Design Principles for Data Spaces," International Data Spaces Association, Berlin, 2021.

corporate entity's capability of being entirely self-determined with regard to its data." [2]. A more pragmatic take is that decentralization supports resilience, lower maintenance costs, technology neutrality, etc. It also supports the way that organisations are using combinations of data from multiple sources to create new insights and services. Anyhow, some decentralized data spaces contain very large, centralized data infrastructures.

Data sharing and exchange are the core of the data spaces. They are made to make data available to individuals, companies, the public sector, and research, i.e. publicly available data. But this is not the same as public data. Data that are available in a data space can either be private or public.

Public data is often data collected, maintained, enriched, and distributed by public institutions. For instance, it could be weather data from the Danish Meteorological Institute (DMI), data about financial performance from the company registration database (CVR), or data about the elevation of the country (the Danish Elevation Model). This is available to everybody, but it does not mean that there are no terms and conditions for the acquisition and use of the data. Even though public data is free to use, there is still a price (paid over national taxes) as well as terms of services (e.g., number of allowed queries over time, etc.).

Private data, on the other hand, is typically collected by a private company or individual, or even by an organisation that is not making the resulting datasets freely available. For instance, this could be some sort of non-personal industrial data from a factory. Here's the data private to a company, but the company can still make it available to other parties, e.g., through a data space. For private data, it is also quite likely that there will be terms and conditions for the acquisition and use of the data, potentially including financial payment.

Another important term from the definition is "**trustworthy**". This is important since the sharing and exchange of data needs to be trustworthy. This involves terms and conditions for the use of data, to which consumers need to adhere. And concerning data sovereignty, the owner of that specific data controls these conditions. Trustworthiness can also apply to the data itself: when providers make their data available, consumers will want the data to be valid, with certain quality criteria. This quality can be enforced through standards for datasets, while the trustworthiness of data exchange can be enforced by secure and standardized protocols.

Trustworthiness between the users will also be increased through a federation that supports security, privacy, and assurance of these in data ecosystems. These ecosystems consist of companies and individuals who share their data through the data space will agree to some common principles. It is described as an ecosystem since it only thrives when many participants engage in sharing and exchanging data. An important observation here is that a culture of using the data space among organisations contributing and using it is just as important as the legal and technical definitions.

3 Commonly agreed on principles

"**Commonly agreed principles**" is another phrase used in the data space definition of IDSA. These principles relate to the governance of data spaces and the policies supporting them. In Europe, there are different initiatives, regulations, directives, and declarations to ensure better interoperability and to make sure that policies are taken into account.

[2] B. Otto, S. Steinbuß, A. Teuscher and S. Lohmann, "Reference Architecture Model," International Data Spaces Association, Berlin, 2019.

3.1 Building blocks – governance, technical

Two main types of building blocks form the data spaces according to IDSA. These are governance building blocks and technical building blocks. The technical building blocks enable the implementation of data spaces i.e., the technical architecture. They are connected to the aspects of trustworthiness and secure data sharing and exchange mentioned in the definition and include APIs, network protocols, and substantially more technical components. The technical building blocks also enable the integration of different systems and platforms.[3]

The governance building blocks are closely related to the commonly agreed principles mentioned in the definition, as these sets are the guiding principles for all implementations of data spaces to ensure data interoperability. They include business, operational, and organizational agreements which set the guiding principles that need to be respected such as decentralization, scalability, and others, in addition to interoperability.[4]

3.2 Minimal Interoperability Mechanisms (MIMs)

Minimal Interoperability Mechanisms (MIMs), is a concept that has originated from the smart city research and collaboration, Open and Agile Smart Cities network (OASC). MIMs are universal tools for achieving interoperability of data, systems, and services between cities and suppliers[5]. As they are based on an inclusive list of baselines and references, MIMs consider the different backgrounds of these organisations and allow interoperability based on a minimal common ground. Implementation can be different as long as crucial interoperability points in any given technical architecture, use the same interoperability mechanisms. The MIMs are vendor-neutral and technology-agnostic, meaning that anybody can use them and integrate them into existing systems and offerings[6]. By implementing MIMs, cities may increase the speed and openness of innovation and development, while also decreasing cost and inefficiency[7]. A challenge of the MIMs and the decentralized structure is that coordination of maintenance is very complex. Thus, the "minimum" in MIMs is essential.

There are currently three MIMs adopted by the 150 member cities of OASC, as seen below:

[3] OPENDEI, "Design Principles for Data Spaces," International Data Spaces Association, Berlin, 2021.
[4] OPENDEI, "Design Principles for Data Spaces," International Data Spaces Association, Berlin, 2021.
[5] https://mims.oascities.org/basics/oasc-mims-introduction
[6] https://oascities.org/wp-content/uploads/2019/03/OASC-MIMs-1.pdf
[7] https://cybersec4europe.eu/associates/open-and-agile-smart-cities-oasc/

MIM	MIM Name	Interoperability Point	Description
1	OASC Context Information Management MIM	Context Information Management API	This API allows to access to real-time context information from different cities.
2	OASC Data Models MIM	Shared Data Models	Guidelines and catalogue of common data models in different verticals to enable interoperability for applications and systems among different cities.
3	OASC Ecosystem Transactions Management MIM	Marketplace API	The Marketplace API exposes functionalities such as catalogue management, ordering management, revenue management, Service Level Agreements (SLA), license management, etc. Complemented by marketplaces for hardware and services.

Table 1: OASC Minimal Interoperability Mechanisms.[8]

Two more MIMS – MIM 4 "Personal Data Management" and MIM 5 "Fair Intelligent AI" – are under development with working groups for each MIM[9], and more MIMs have been defined, namely the MIM 6 "Security Management", MIM 7 "Geospatial Information Management", MIM 8 "Ecosystem Indicator Management", MIM 9 "Data Analytics Management" and MIM 10 "Resource Impact Assessment"[10]. The practical applications of the MIMs can be found in the catalogues from OASC[11] and City-by-City[12], wherein addition, there is a more in-depth description of each MIM, how they are proposed and governed, information about reference implementations, and more on OASC's MIMs Gitbook page[13].

OASC members **f**ormally adopt, endorse, support, and advance the Minimal Interoperability Mechanisms and their underlying principles of open data; open APIs; open access; and 'driven by implementation'. The proposal for a new MIM is prepared by initially one city or partner who then opens the process for consideration and enrichment by other members and partners in the Council of Cities (CoC), representing all OASC partners and working groups. It then goes through a process, where it is built, tested, shared, and more until it has 'gone live' and will be maintained, evolved, and promoted by the appropriate bodies of OASC[14].

3.3 Living-in.EU (MIMs+)

The Living-in.EU (LI.EU) declaration and initiative ensure that local priorities are considered when scaling digital solutions. The LI.EU technical specifications (called MIM+) include EU policies such as INSPIRE, CEF, and EIF; it refers to established standards from the SDOs such as ISO, OGC, ETSI, where available (such as geospatial, food, farming, energy, water, etc.). A challenge is that data harmonisation, which ensures that data models can be harmonised with shared data models and between different standards, based on experience from Denmark, is time-consuming before aligning all these standards for instance in the Danish Basic data program[15]. MIM+ also aligns with the

[8] https://oascities.org/wp-content/uploads/2019/03/OASC-MIMs-1.pdf (page 2)
[9] https://oascities.org/minimal-interoperability-mechanisms/
[10] https://mims.oascities.org/basics/oasc-mims-introduction
[11] https://oascities.org/list-of-cities/
[12] https://catalogue.city
[13] https://mims.oascities.org/
[14] https://mims.oascities.org/basics/oasc-mims-governance
[15] https://datafordeler.dk/media/1128/grunddata-modelregler.pdf

emerging UN SDG framework in U4SSC, which includes both indicators and architectures. It also links to, other global regions and their markets which are important for the EU, this includes Japan and India. Backed by suppliers in Europe and elsewhere, including the well-known hyper scalers, but with distinct terms and conditions in line with EU priorities and policies. In short, the development and evolution of Minimal Interoperability Mechanisms in Living-in.EU (MIMs Plus) is a way to help both the demand side, suppliers, and regulators invest in capabilities that solve the problems at hand, locally. Catalogues of certified services with easy-to-procure procedures, accompanying legal templates and financial packages are emerging based on this approach[16].

LI.EU has created an architecture framework model, organised according to the MIMs, as seen below:

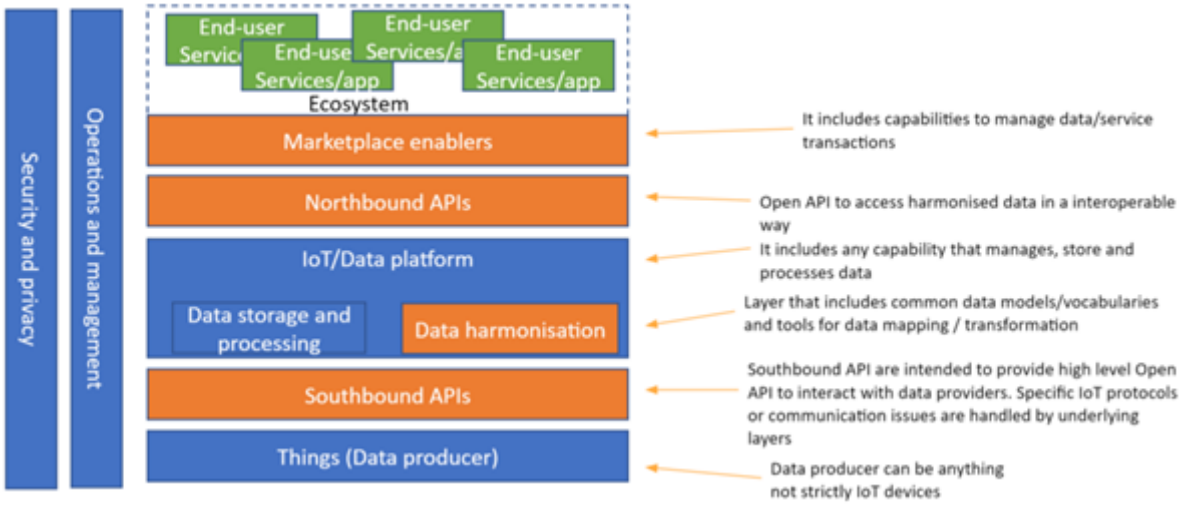


Figure 1: High-level architecture framework model.

It provides a description and guidelines of a common architecture/framework, including a layered overview positioning of all the components and interfaces, as well as the associated requirements and specifications[17]. There are several other architecture frameworks from other organizations such as Inspire and Datafordeleren.

3.4 Roles

In the state-of-the-art investigations on data spaces, there are many roles concerning these data spaces. In this white paper, three main roles will be used: *Consumer, Provider, and Leadership (performing Governance)*. The *consumer* is an organisation, company, or individual who wants to use the data space. The consumer will gain access to certain datasets by accepting terms and conditions for acquisition and use of the given data as well as potential financial payment. The *provider* is an organisation, company, or individual who creates or collects, and distributes data through the data space. The *governance* role is a bit different from the two others since this one will be more "in charge" of the data space. The role involves legislation or principles around the use of data spaces, but not controlling the data since that belongs to the owner of the data. The role also involves the trust and security of the data space, i.e., providing regulations for the use of data spaces, who is certified to use

[16] https://www.living-in.eu/sites/default/files/files/Consolidated-Report-on-Tech-Specs-v2.pdf
[17] https://www.living-in.eu/sites/default/files/files/Consolidated-Report-on-Tech-Specs-v2.pdf (figure from page 3)

it, as well as making the data space secure from a technical point of view. For example, that it isn't possible to bypass the regulations and gain unauthorized access to data.

3.5 FAIR data principles

In relation to the generation of data for gaining insights and generating value, it is worth mentioning the FAIR guiding principles. These principles can be originated for researchers and scientific data but are also relevant from the perspective of data spaces, hence they have been foundational for the INSPIRE directive, OGC, ISO, and many more. The principles are developed for improving data management and stewardship to ensure proper collection, annotation, and archiving, as well as long-term care of digital assets. FAIR stands for *Findability*, *Accessibility*, *Interoperability*, and *Reusability* and these are the four principles meant for guiding data producers and publishers. Even though the principles are not developed specifically for data spaces they still make a lot of sense in our context. Data spaces are meant to be a place for trustworthy data sharing and exchange, which means it is a place to go to for sharing your data and getting access to others' data. This fits well with the findability and accessibility of the FAIR principles. Findability and accessibility are very important aspects of data spaces because when an individual or company wants to start working on a new project requiring data, the first question is how to begin looking for data? The answer will then be the data spaces or more specifically the data space made for the given domain. Interoperability has already been mentioned and is also an important part of data spaces. The data must be reusable by all the different users getting access to it. This can for example be ensured through the commonly agreed principles on how to use the data as well as through clearly described data definitions and rich metadata.[18]

3.6 Reusability

Another important aspect of data is reusability, which commonly agreed principles could help achieve. A dataset having high reusability means, for example, high quality of data and high user-friendliness; or in other words that data has a low number of errors and can easily be understood and used by others.[19,20] Some of these principles could for example be public standard descriptions of datasets and data catalogues. Concerning this, the Danish Agency for Digitisation has developed such a standard description based on an analysis of several companies and organizations' needs for metadata.[21] Furthermore, the Danish Agency for Digitization has developed a "common language for data quality" which is not a tool for ensuring data quality but a step towards better communication about data quality between the data provider and data consumer.[22]

4 First-generation data spaces and their challenges

As mentioned earlier, data spaces are, in this white paper, attended as the newest evolution of policies and technical tools for data exchange. Therefore "first generation of data spaces" includes the simpler/earlier concepts such as data catalogues, data portals, data distribution platforms, or data hubs. These are all different kinds of concepts but evolutions of data spaces, as described earlier. The simpler concepts such as data catalogues are typically implemented as a database with a collection of all the data sets available about the topic. Basic functionality for searching for data, but no structures

[18] M. Wilkinson et al., "Comment: The FAIR Guiding Principles for scientific data management and stewardship", 2016, *Scientific Data*
[19] https://arkitektur.digst.dk/metoder/faelles-sprog-datakvalitet
[20] https://arkitektur.digst.dk/metoder/vejledning-i-genbrug-af-data-i-selvbetjeningsloesninger
[21] https://arkitektur.digst.dk/rammearkitektur/datastandarder/dcat-ap-dk
[22] https://arkitektur.digst.dk/metoder/faelles-sprog-datakvalitet

or formats predefined, and limited exportability. One example could be Copenhagen city data exchange which can be described as a data hub – a more centralized data space.

Currently, several data services exist in Denmark and internationally although the term commonly used to describe them might not be precise "Data space", but instead data portal, data catalogue, or data distribution platform. These could be viewed as the first generation of data spaces where different organizations are gathering data or making their data available through online services. There is often not one specific place where the data can be found so instead the consumer would have to research a different type of data distribution platforms or catalogues to see what kind of data they may use. Because they are so scattered, it makes it difficult for consumers such as SMEs to find usable data they can use in different applications to generate additional value. If the data was gathered in one “data space” this might be the only place to look for relevant data. Thus, saving a lot of time and work researching looking for the right data. This does not necessarily have to do with the data being public. It is rather a historic development deriving from different domains providing data for domain-specific use. Now that data has become a cross-sector and cross-institutional asset, the heterogeneous distribution landscape has become an obstacle for many users wanting data from several domains and distribution channels.

In this white paper, we define a catalogue of different Danish and international data catalogues and portals as they might be viewed as the first generation of data spaces, to provide the reader with an idea of where to start searching for data useable in their specific applications.

The European Commission has made a strategy for data, which aim is to create a single market for data described in [3]. This is to ensure Europe’s global competitiveness and data sovereignty. The plan is to establish nine common data spaces, each for different domains. The domains are:

- Industrial (manufacturing) data space
- Green Deal data space
- Mobility data space
- Health data space
- Financial data space
- Energy data space
- Agricultural data space
- Data spaces for public administration
- Skills data space

Since an important application domain for this white paper is data for smart cities, it is worth mentioning that the smart city domain spans several of these sectors. For instance, energy management is important in a smart city, mobility includes transport which is also worth mentioning, and so on.

Some of the first generations of data spaces in Denmark are also related to smart cities, where municipalities have made their data open. This data includes for example traffic data, parking data, physical infrastructure, and more.

As mentioned before, many of these data spaces are from municipalities and other public organizations. There are some data spaces such as opendata.dk which collects the datasets from these different organizations and makes them available through their portal. But in general, the main challenge of the first generation of data catalogues is how scattered they are. Opendata.dk does not

include all open data in Denmark, so if consumers are looking for data, they must also look around and find the other data catalogues to see which one(s) they need.

Other challenges of the first generation of data spaces include the lack of structure as they can be seen more like a collection of datasets, or data catalogues, instead of systems with higher interoperability and possibilities for integration. Some might have APIs and others might be more like data catalogues. Data catalogues are mostly collections of datasets (data files). Also, they do not necessarily have clear agreements on data age or update frequency. This results in potentially old datasets or datasets that are rarely updated. But it also depends on the source of the data, i.e. the organization creating the data, as well as how often new data is created. A first-generation data space could for example be a simple data catalogue with datasets gathered from different sources. But it could also be a more complex data space with an API for querying data. If there is such an API, it is also more likely that the update frequency is higher as the users query the data directly.

A rather consumer-driven and almost anarchistic approach to data spaces is the use of web scraping which is, the practice of harvesting data, sometimes illegally, from websites. One could argue that this is a community that consumes data from a decentralized infrastructure.

EU´s vision for Digital Europe will be carried out via the Data strategy that points to common Data Spaces as a tool to increase data access and data use. Individuals and organizations will have control of their data and share it through these data spaces. We believe that the data spaces available 5 years from now will be a game-changer for how we use data in our businesses.

5 Process of working with data spaces

There are several types of processes, which involve working with data spaces, with different starting points. For instance, some users might know the precise data they need, and only need a process to retrieve and consume it. While some other users might only have a relevant case and may not even be sure that the relevant data is available to them. Despite those differences, we can extract some general process steps. It should be noted that these experiences are based on "first-generation data spaces" and how to practically work with data in general today. When data spaces are properly launched the process will hopefully be simpler.

First, scoping the problem is essential to be able to address some precise questions to oneself as well as to another domain expects. The precise questions are needed to be able to precisely define the needs, such as the type of information that is demanded, its precision & granularity, freshness, pricing, etc. without overestimating the requirements. Indeed, oftentimes a relatively low-quality dataset is sufficient to get started. An essential aspect of the first step is: Based on the question(s) we are trying to answer, what data can directly or indirectly enhance our knowledge to answer the question and where do we find it.

Then comes the quest for identifying the relevant data sources. Here the data portals can be especially useful (see for example section 9.2 about CKAN) as well as lists of such instances (see section 8). Just browsing the available datasets can also generate new ideas and improve the initial problem statement.

It is not always trivial to imagine that a dataset exists, so in addition to the data portals, looking at the state-of-the-art such as related solutions, e.g., from governments, can put you on the right track. Furthermore, if the ideal dataset cannot be found, maybe combining two or more related datasets can generate the desired information.

While data spaces are a great help for the problems above, contacting a domain expert and/or those generating related datasets is still invaluable, especially because there is often a significant amount of tacit knowledge that is not easily discoverable online. This is especially true for restricted datasets, which might require a contract or NDA to access. Finding the right person in the organisation is not always straightforward, though, and having a clear problem statement helps there as well. Hopefully, the fully implemented dataspaces will contain metadata, that resolves these challenges.

When some candidate datasets have been identified at last, then the consumer needs to become familiar with them. While some data spaces readily expose ways to browse or toy with the data, dedicated tools are needed (see also section 9) before proceeding further.

Before kicking off a full-fledged implementation project based on the newly found datasets, we recommend a step with a proof of concept consuming the said datasets. Indeed, during this phase, several hidden issues might surface, such as a gap between the actual data quality and what is advertised, e.g., in terms of data freshness, uptime, throttling (limiting the number of requests), etc.

Hereafter the process of analysing the data through data science begins, but that is another challenge for another whitepaper.

6 Policy landscape

The European Commission's European Strategy for Data[23] aims at creating a single market for European data. Common European data spaces will ensure that more data becomes available for use. They will give access to personal as well as non-personal data, in a secure manner according to GDPR, ensuring growth and creating value for businesses in a manner where the carbon and environmental footprint is minimised. EU law will be enforced in this market for data spaces, where all products and services will comply with the relevant regulations of the EU single market (4-5). The data may then flow within the EU and across sectors, with fair rules and full respect of personal data protection, consumer protection legislation, competition law, etc. Of course, the creation of such a space depends on the investment in next-generation technologies and infrastructures, as well as in digital competencies such as digital literacy.

To ensure the creation of such spaces, the European Commission is therefore investing in a data infrastructure of the future, with a focus on edge computing, high-performance computing/quantum computing, cyber-security, low-power processors, and 6G networks[23]. These investments are vital to equip Europe with the right infrastructures, computing power, encryption capacity, and cybersecurity tools to process data. As part of these investments, the European Commission is investing in a High Impact project on European data spaces and federated cloud infrastructures. The project will address the specific needs of industries in the EU, including hybrid cloud deployment models that allow data processing at the edge with no latency (in the cloud-edge continuum), through the funding of infrastructures, data-sharing tools, architectures, and governance mechanisms for data-sharing and AI ecosystems. The project will support and benefit the European ecosystem of data-intensive companies and support them and the public sector in their digital transformation.

The European Commission has proposed a Regulation on European data governance[24] as part of its data strategy; the Regulation will play a vital role in ensuring the EU's leadership in the global data

[23] The European Commission. Communication: A European Strategy for Data. 2020. https://ec.europa.eu/info/sites/default/files/communication-european-strategy-data-19feb2020_en.pdf
[24] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767

economy and will empower users to stay in control of their data and encourage the creation of common European data spaces in crucial sectors. These sectors include health, the environment, energy, agriculture, mobility, finance, manufacturing, public administration, and skills[25].

This is seen in the EU funding programme for research and innovation, the Horizon Europe programme, where it is focusing on data spaces within e.g., the European Green Deal where there is a call for "Common European Green Deal data space to provide more accessible and exploitable environmental observation data in support of the European Green Deal priority actions"[26]. The new EU funding programme and the Digital Europe programme[27], also center data spaces. The programme is focused on bringing digital technology to businesses, citizens, and public administrations. As part of their actions within the realm of artificial intelligence, they wish to focus on the deployment of EU-wide common data spaces based on a cloud-to-edge federated infrastructure and promote testing and adoption of Artificial Intelligence-based solutions[28].

European data spaces are also being supported by other projects, such as GAIA-X. GAIA-X is a project initiated by Europe for Europe and beyond. It aims to develop common requirements for a European data infrastructure. With GAIA-X, representatives from business, science, and politics on a European level create a proposal for the next generation of a European data infrastructure: a secure, federated system that meets the highest standards of digital sovereignty while promoting innovation[29].

In short, many different actors are working on data spaces and so we should expect this subject area to be evolving at a steady pace.

6.1 GAIA-X and IDSA

There are currently two European Data Spaces initiatives that have gained traction, namely GAIA-X and IDSA.

GAIA-X is a European project, based on European values, their intention is both to connect existing cloud services and spark innovative new modes of connectivity to create a federated digital infrastructure for Europe[30]. GAIA-X support the goals of the European data strategy, published by the EU Commission in 2020[31].

The objectives that are stated as part of the Commission's specific measures, such as European data spaces, federated cloud, a rulebook, cloud service marketplace, are likewise integrated in the GAIA-X project. In GAIA-X, representatives from business, politics, and science from Europe and around the globe are working together to create a federated and secure data infrastructure. Companies and citizens will collate and share data – in such a way that they keep control over them. They should decide what happens to their data, where it is stored, and always retain data sovereignty[32].

GAIA-X wants to create a data structure, a networked system, that links up various suppliers of cloud services through interoperable data exchange. It will also act as a repository that may be used to

[25] EC's data strategy: https://digital-strategy.ec.europa.eu/en/policies/strategy-data
[26] Horizon Europe call: https://www.efsa.europa.eu/en/funding-calls/common-european-green-deal-data-space-provide-more-accessible-and-exploitable
[27] The Digital Europe Programme: https://digital-strategy.ec.europa.eu/en/activities/digital-programme
[28] Digital Europe Programme actions: https://ec.europa.eu/info/strategy/eu-budget/performance-and-reporting/programmes-performance/digital-europe-programme-performance_en
[29] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html
[30] https://internationaldataspaces.org/we/gaia-x/
[31] https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-a-pitch-towards-europe.html
[32] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

search for specific data services. Gaia-X aims to reduce Europe's reliance on international cloud giants. Instead, businesses should look to European solutions, protected by European data laws[33].

GAIA-X strives to have data spaces for the following domains: energy, finance, health, industry 4.0/SMEs, mobility, public sector, smart living, and agriculture. GAIA-X aims to help to break up data silos and avoid data lock-ins[34], by e.g., linking existing European cloud services and making it easier for businesses to exchange data access the various domains[35].

The International Data Spaces Association (IDSA), a non-profit organisation, is an integral part of GAIA-X as GAIA-X uses the International Data Spaces standard (IDS) in its architecture[36]. The IDS standard is the blueprint for data exchange itself, based on European values, such as data protection and security, equal opportunities through a federated design, and the guarantee of data sovereignty for the creator of the data and trust between participants. A critical success factor is to ensure data sovereignty and interoperability — a shared goal of GAIA-X and IDSA, which is why IDSA concepts are an integral element of the GAIA-X architecture[37].

The International Data Spaces (IDS) initiative aims at cross-sectoral data sovereignty and data interoperability, and so its Reference Architecture Model (IDS-RAM) is based on open standards and contributes to global standards. By specifying data usage constraints, it defines the terms and conditions for the data economy and by describing an open software architecture and publishing opensource software codes, it ensures maximum adoption. The IDSA currently has more than 110 members from more than 20 different countries. The IDSA argues that if IDS-RAM is used as a blueprint for the European Data Space, it may speed up implementation in key areas[38].

7 Key actors in Denmark

Several actors in Denmark have created data catalogues, (which could be viewed as the first step towards data spaces), such as Open Data DK, Datafordeleren, Digitaliseringsstyrelsen, Dataforsyningen, Virk, Folketinget, Statistikbanken, and Miljøportalen.

Open Data DK is an association of Danish municipalities and regions that since 2016 have collaborated to open their data, amongst others on a common open data portal. Everyone can use the data in the data portal for free without registration. All municipalities and regions can use the portal to publish their data sets. The data portal is based on the Open Source software CKAN from the Open Knowledge Foundation[39]. Open Data DK is funded by the Danish Business Authority (Erhvervsstyrelsen) and Local Government Denmark (KL) and together with other actors, they are currently also part of a public project regarding how we should deal with open data in the future. Furthermore, Open Data DK is harvested by the European open data platform automatically through structured metadata and it is evolving to fit emerging trends, for example by being able to support real-time IoT data from the OS2iot-platform.

[33] https://www.techrepublic.com/article/what-is-gaia-x-a-guide-to-europes-cloud-computing-fight-back-plan/
[34] https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-a-pitch-towards-europe.html, pages 3, 10-11
[35] https://www.techrepublic.com/article/what-is-gaia-x-a-guide-to-europes-cloud-computing-fight-back-plan/
[36] https://internationaldataspaces.org/
[37] https://internationaldataspaces.org/we/gaia-x/
[38] https://internationaldataspaces.org/wp-content/uploads/IDSA-Position-Paper-Implementing-European-Data-Strategy-Role-of-IDS1.pdf
[39] https://www.opendata.dk/about

The Danish Agency for Data Supply and Efficiency (Styrelsen for Dataforsyning og Effektivisering) is responsible for the Data Distributor (Datafordeleren). The Data Distributor is the distribution platform, which makes basic data from several authorities available in one place. It shows basic data from several authorities and ensures that authorities, companies, NGOs, citizens, and more have easy and secure access to basic data in one single access point.

The Data Supply (Dataforsyningen) is a single point of access to mainly spatial data and maps. Agency for Data Supply and Efficiency provides an overview, information, case stories, and access to a variety of data products deriving from different sources and data providers.

The Danish Agency for Digitization has created a data catalogue, a so-called joint public data set catalog that provides an overview, of which public data sets are available, where they are located and whether they are accessible. The dataset catalogue contains only metadata, i.e. a description of the dataset[40].

Statistics Denmark also has a data catalogue called StatBank Denmark. It contains uncommented numbers collected by Statistics Denmark from the Internet. StatBank Denmark can offer more possibilities of exploring data than the printed publications from Statistics Denmark[41].

The Danish Environmental Portal is a data portal that offers a range of digital solutions that give access to data on the environment, including among other things web services, a user administration system, and access to legacy archives with data from the old counties. With the creation of the Danish Environmental Portal, private citizens and professionals alike, can draw on and update the State, County, and municipal databases. At the same time, the cooperation between all levels of government with regards to data and digital government lays a foundation for the effective handling of environmental issues in Denmark[42]. It is possible to find links to these data spaces and more below in the examples section.

Many SMEs need to build internal knowledge on data spaces, data catalogues, and data exploration, and actors to help them within the process. This would be a great advantage to SMEs. An essential point to mention about such actors is the lack of a specific group, namely private data brokering companies. It is very hard to find those, who can help companies find the right data sources, perhaps due to the lack of monetary gains for the brokering companies in providing such a service.

8 Examples of Data Spaces and how to find them

As mentioned earlier one might view data catalogues, data portals, and distribution platforms, as early versions or a first-generation of data spaces. This section gives different examples of data spaces both in Denmark and internationally. As these data spaces are already established it is worth mentioning that they are in the first generation of data spaces. Thus, they might be lacking in several aspects regarding interoperability and integration with platforms. Another term to describe some of them could therefore be data catalogues. Data catalogues are simply collections of data sets where one can find data sets and download them, for example in the form of CSV or Excel files or access them through APIs. This means there is not much if any, integration with user systems. There might also not be any agreement or schedule for updating datasets, which means they could potentially be old and the latest

[40] https://datasets.catalogue.data.gov.dk/
[41] https://www.statistikbanken.dk/statbank5a/default.asp?w=1536 (under "hjælp" and "Statistikbanken generelt"
[42] https://miljoeportal.dk/english/

update to a given dataset might also be the last. However, it is worth looking into the metadata related to the data, concerning the use of that data, for instance through the catalogue services.

On the other hand, the vision of data spaces described at the beginning of this white paper is more complex. They are, from a data space vision perspective under development but will have more possibilities regarding interoperability and integration between systems. Anyhow, they are very mature data catalogues that are very close to data spaces and have been operational for more than 20 years. This is both because the data space vision is targeting much wider on an international scale, and because there are going to be these commonly agreed principles on how to use and operate the data spaces.

Table 2lists examples of current Danish data catalogues/data portals:

Data space name	Link
Open Data	https://www.opendata.dk
Datafordeleren	https://datafordeler.dk
Digitaliseringsstyrelsens fælles offentlige datakatalog	https://datasets.catalogue.data.gov.dk/
Dataforsyningen	https://dataforsyningen.dk
Virk Data Datakatalog	https://data.virk.dk/datakatalog
Folketingets åbne data	https://www.ft.dk/da/dokumenter/aabne_data
Statistikbanken (Danmarks Statistik)	https://statistikbanken.dk/
Danmarks Miljøportal	https://miljoeportal.dk/

Table 2: *Examples of current data spaces in Denmark.*

Table 3 gives examples of current international data spaces/data catalogues:

Data space name	Note	Link
Data.europa.eu	Includes data from Denmark	https://data.europa.eu/da
The World Bank	Includes data from Denmark	https://data.worldbank.org/
Data.gov	U.S. government data – using CKAN	https://www.data.gov/
Open.canada.ca	Canadian government data – using CKAN	https://open.canada.ca/en/open-data
Global open data index (Open Knowledge Foundation)	Includes data from Denmark	https://index.okfn.org/
Datarade		https://datarade.ai/search
OpenDataSoft		https://data.opendatasoft.com
Kaggle	Also includes tools for working with datasets	https://www.kaggle.com/

Table 3: *Examples of international data spaces. Some include data from Denmark.*

Table 4 gives examples of other data space-related services. It is not possible to get data from these but instead, they can help locate the data spaces where you can get the data you need. Thus, they can be described as a layer above the data spaces, helping find the right data space or catalogue.

Name	Note	Link
Open Data Inception	Hub for data spaces	https://opendatainception.io
Open Data Monitor	Hub for data spaces	https://opendatamonitor.eu
CKAN	Open-source software for open data catalogues	https://ckan.org/

Data distributor catalogue	List of data distributors and links to them	https://arkitektur.digst.dk/kataloger/datadistributoerkatalog

Table 4: Examples of other data space-related services.

9 Tools for data spaces

While the main substance of the data spaces is the data itself, some practical software tools are needed to expose the dataset efficiently, as well as to help create and administrate data spaces.

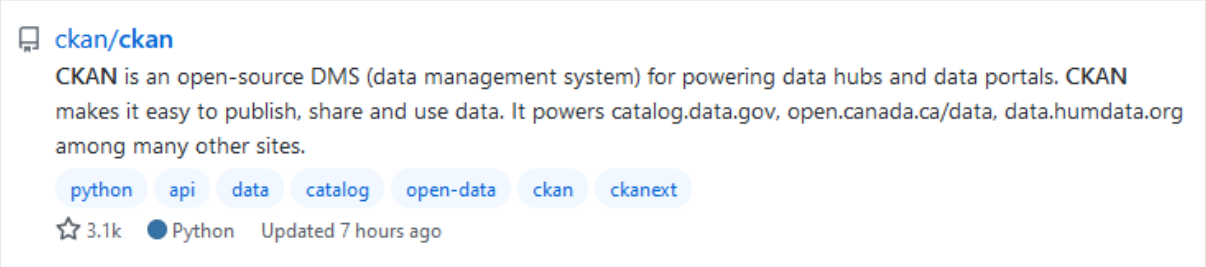
9.1 Knowledge graphs

When attempting to formalise datasets publication, discovery, exchange, and interrelations, many academic efforts have been cast at the highest quality level of open data, namely linked semantics data[43].This is in addition to standardised vocabularies (ontologies), formats (e.g., RDF), and compatibility with related technologies (schemas, semantic query languages, etc.). Those are forming knowledge graphs. Popular generic examples include DBpedia[44] and later Wikidata[45].

However, the vast majority of data, also of open data, is conveyed through less powerful formats such as tabular data, or XML or JSON data representations (serialisations) that are not using any formal vocabulary (ontology). This reinforces the need for more powerful data management and data-sharing software platforms.

9.2 CKAN

One of the most known Open Source software tools since the first generation of data spaces is the data management system CKAN[46]. Today, it is serving as the main tool for data portals of several national governments (Canada, Singapore, USA, Australia…), cities around the world (Berlin, Helsinki…), and enterprises (ESO, HDX…). CKAN is used in this whitepaper as a practical example of how to work with data, but many others exist. In particular, it is worth mentioning GeoNetwork[47].



In Denmark, CKAN is notably used for Open Data DK[48] with public data from several municipalities (e.g., Aarhus, Copenhagen) and authorities (e.g., Vejdirektoratet), and by LEGO (as an internal data hub).

In addition, to allow searching for datasets based on many metadata criteria, CKAN also offers some basic data visualisation.

[43] https://5stardata.info
[44] https://www.dbpedia.org
[45] https://www.wikidata.org
[46] https://ckan.org
[47] https://geonetwork-opensource.org/
[48] https://www.opendata.dk

9.3 Alternative data sharing tools

There are a few other closed-source competitors on the market, most noticeably the data-sharing platform OpenDataSoft[49], used by cities such as Paris and Vancouver, includes some data visualisation and analysis features.

Finally, GeoNetwork[50] is an open-source catalogue application to manage spatially referenced resources.

9.4 Data exploration and data science tools

The relatively simple data visualisation and analysis feature built-in the above-mentioned data platforms are a good start to get a first sense of what a given dataset looks like. It is indeed important to get a good understanding of the properties of the dataset before using it. But to do that properly, loading at least a part of the dataset in a dedicated tool is essential.

Depending on the user's background, different categories of software are commonplace, for instance:

- For a (business) analytics specialist, the first move would be to load the dataset in a tool such as Microsoft PowerBI or Tableau, allowing many types of visualisations.
- For a statistician, it would be loading the dataset in her favourite environment such as R, to perform descriptive statistics, regression, visualisations, etc. If the background is social science, then the statistical package would often be SPSS.
- For an engineer, it can be loading the dataset in a numeric computation environment such as MATLAB or GNU Octave, allowing the application of various mathematical tools to the data with the corresponding visualisations.
- For a database specialist (e.g., DBA), it would be loading the dataset in an SQL database such as PostgreSQL to perform various queries, like counting occurrences, grouping, clustering, joining with other datasets, etc. For the example of PostgreSQL, the database can be supplemented with time-series features thanks to the Timescale extension, and with spatial & geography thanks to PostGIS.
- For a geospatial data specialist, it would be loading the dataset into a Geographic Information System such as QGIS (potentially after the database step with e.g. PostGIS) or Esri ArcGIS to visualise and navigate the data on a map.
- For a data science specialist, it would be loading the dataset in an environment such as Jupyter Notebook, from which many (typically Python) libraries can be taken advantage of to analyse and visualise the data before proceeding with some machine learning (ML) modeling.

All the above-mentioned tools are valid approaches and using more than one of them is desirable. They are all able to load at least basic datasets in the most common standard formats. In the IoT world, where datasets consist mainly of time-series (repeated measurements over time), some software components are quite commonplace, such as:

- For data ingestion and pre-processing: Node-RED (IBM), Apache Nifi, etc.
- As a database: TimescaleDB (based on PosgreSQL), Prometheus, Elasticsearch, InfluxDB, etc.
- For analytics / visualisation: Grafana, Kibana, etc.

[49] https://www.opendatasoft.com
[50] https://geonetwork-opensource.org

9.5 Related integrated platforms

It is also worth looking at platforms that are not IoT-focussed, as they might have some inspiring characteristics.

For instance, Kaggle[51] (Google) is an online community of data scientists and machine learning practitioners allowing to publish and search for datasets here especially for machine learning purposes. In addition to public datasets, the platform also offers the ability to host (snippets/kernels) and run (Jupyter Notebooks) software code directly on it, with access to hardware computing resources (CPU, GPU). Finally, there is a strong community activity with competitions around concrete problems.

Another example is WolframAlpha[52], which is a computational knowledge engine, which in addition to advanced mathematics capabilities, incorporates a vast knowledge base. It allows complex queries – which can be expressed in natural language – about real-world aspects including science, technology, society, culture, personal life, etc. It is more of an example of taking advantage of public datasets with their own added value. Other examples of data semantics that structure data in knowledge graphs are Wikipedia and Google Rank brain.

[51] https://www.kaggle.com
[52] https://www.wolframalpha.com