data·pitch
INNOVATION PROGRAMME

DATA SHARING TOOLKIT

Lessons learned, resources and recommendations for sharing data

Gefion Thuermer, Johanna Walker, Elena Simperl

TABLE OF CONTENTS

List of case studies

Key resources

ABOUT THIS TOOLKIT

Data plays a major role in the European economy, and building a European data economy is one of the strategic goals of the European Commission. Through the increase of data science techniques, not least Machine Learning (ML) and Artificial Intelligence (AI), the value and role of data as an asset becomes ever more crucial. This has made it more important for data to be accessible. However, much of the data that many solutions require are held within private organisations - and are only available if they are *shared*. Data sharing in this sense means *allowing third parties specifically permissioned access to datasets to generate value.*

This toolkit has been developed to help organisations that want to generate value by sharing data or facilitating data sharing. We explain the concept, challenges, and processes to enable successful data sharing, and provide resources and recommendations. It is derived from experience collected in the Data Pitch programme and related national and international initiatives, such as the Smart Cities Innovation Framework Implementation (SciFi), the European Data Incubator (EDI), as well as several recent pilots for data trusts in the UK.

The document is structured in three main sections:

1. An introduction to the fundamental notions of data and data sharing, including a discussion of the benefits of data sharing;
2. An introduction to Data Pitch, the three year data sharing programme that commissioned this research;
3. Recommendations and resources for data sharing, leveraging feedback and experiences from interviews and case studies carried out in the context of Data Pitch and other related works.

The toolkit offers practical advice and guidance to:

- organisations which have data that they want to share in order to investigate its value, such as corporations;
- organisations which wish to facilitate data sharing, such as data marketplaces;
- organisations which use artificial intelligence and machine learning and wish to enter data sharing agreements with data holding organisations; and,
- individuals who want to enable data sharing on a practical level, such as innovation managers.

Most research on which this toolkit was focused on open innovation for new products and services in the commercial market. However the toolkit can be equally used to guide data sharing for social good, knowledge development or a multitude of other aims.

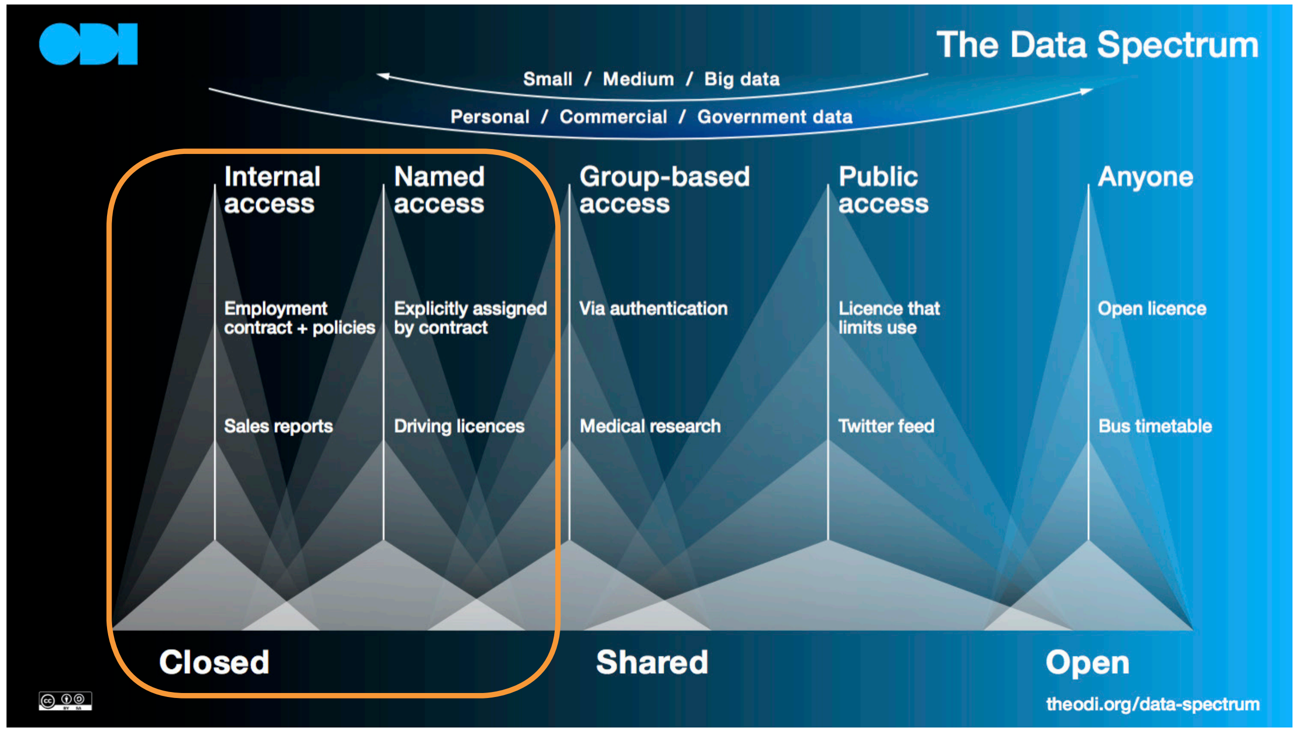
How to use this toolkit

1. Learn about data sharing in section one, to understand the relevant concepts.
2. Read about an example in section two, to see how these concepts can be applied.
3. Follow the steps and use the resources in section three, to help you implement your own data sharing.

ABOUT DATA SHARING

Data sharing means allowing third parties specifically permissioned access to datasets to generate value.

Governments, organisations, public bodies, third sector organisations etc., hold vast amounts of data. Some of this data can and should be made openly available; other data, for competitive or privacy reasons, very much should not. However, in between is a wide range of possible modes of data availability. Data can therefore be mapped on a spectrum, such as that created by the Open Data Institute (below) in which they differentiate between these various modes of data access via the legal basis for access.


The Data Spectrum

Small / Medium / Big data

Personal / Commercial / Government data

Internal access — Employment contract + policies — Sales reports

Named access — Explicitly assigned by contract — Driving licences

Group-based access — Via authentication — Medical research

Public access — Licence that limits use — Twitter feed

Anyone — Open licence — Bus timetable

Closed

Shared

Open

theodi.org/data-spectrum

On one end of this spectrum is open data, publicly available for anyone to use, and often open government data, through data that is shared amongst specialist groups, such as research data, to data that is currently closed, such as stock lists, or sales reports. In this toolkit, therefore, we define 'data sharing' as the sharing of otherwise closed data within or between organisations. The impetus for this may vary. In this toolkit, we address how data of many types at the closed end of the spectrum can be selectively shared to create value for both the data holder and user.

Why is data sharing important?

Currently closed data is considered to have a major role to play in contributing to the European data economy, which is projected to be worth €739bn by 2020.[1] This is why building a European data economy is one of the strategic goals of the European Commission, which named data driven innovation as "a key enabler of growth and jobs in Europe."[2]

With the increased use of Machine Learning (ML) and Artificial Intelligence (AI), the value and role of data as an asset has increased as well. These new technologies can help to tackle many modern social challenges, but are dependent on the availability of large amounts of data. While a multitude of open data is freely available, this is often of limited quality, unstructured, or inconsistent.[3]

Many solutions require types of data which are held within organisations, and not intended for public consumption.[4] This data could be privately held data that is of public interest, such as transaction data from mobile telecom operators, sensor data from personal communication devices or from smart electricity consumption meters (sometimes known as 'business to government' or 'b2g' data sharing).[5] It could be privately held data that could be of economic interest, and used by skilled technology companies to develop innovative new products, services and markets. In this way the European Commission envisages private sector data as a key driver of innovation and competitiveness in Europe.[6] Similarly, the OECD considers data sharing as "an effective means through which the social and economic value of data can be maximised."[7]

Further, the data need not necessarily be complete, static data sets; it may equally be metadata or synthetic samples, and with the growth of the Internet of Things, is likely to be streamed sensor data.

Making this data available for specified purposes can unlock value for the organisation that holds it, for data users working with this data, or for the general public. This is why data sharing is crucial.[8]

[1] European Commission, 2019a
[2] European Commission, 2019b
[3] Perez, 2018
[4] Verhulst & Sangokoya, 2015
[5] European Statistical System, 2017
[6] European Commission, 2018
[7] OECD, 2017
[8] European Commission, 2019b

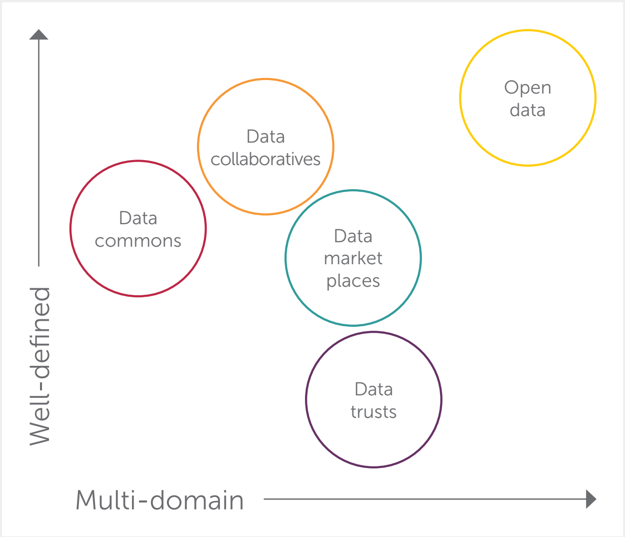
Forms of data sharing

The value of data may not be maximised with one single instance of data sharing. Rather, sharing data unlocks so-called big data value chains,[9] where large amounts of data are collated, processed or transformed in several related steps.

This requires established frameworks in which data can be shared, not only once, but consistently. Walker et al. (2019) identify the following list of already established practices that enable a form of data sharing in different contexts:

- **Data commons:** Resources are held in common, accessible to all members of a group. This primarily occurs for medical and related interoperable data (and tools) between researchers,[10] but also in the energy sector.[11]
- **Data collaboratives:** Private data which benefits society and the environment is shared for social good.[12]
- **Data marketplaces:** Intermediary platforms or online stores through which data can be bought or sold.[13]
- **Data trusts:** There is no one definition of what a data trust is (yet). As a working definition, O'Hara suggests that data trusts work within the law to provide **ethical**, **architectural** and **governance** support for trustworthy data processing.[14] Data trusts can be for internal use only[15] or to facilitate sharing externally, to support AI innovation[16] or social good,[17] or to protect citizens.[18]
- **Open data:** Data that is licensed and available for anyone to access, use and share for any purpose. Personal data can never be open data.

Data commons and data collaboratives are relatively well-defined concepts with specific aims. Data trusts and data marketplaces are more fluidly defined but are appropriate to a wider range of industry sectors and aims.



Case study: Dawex, a data marketplace

Dawex is a leading data marketplace, which allows organisations to make their data available for purchase under licence. They believe marketplaces will accelerate data sharing, because they make it scalable and affordable.

Like all markets, Dawex brings liquidity - it provides a forum for matching offer and demand. In fact they see themselves as a kind of 'AirBnB for Data'. Like AirBnB's rooms, some data sets are not sold at all and some are very popular. Also like AirBnB, Dawex stimulates supply and demand.

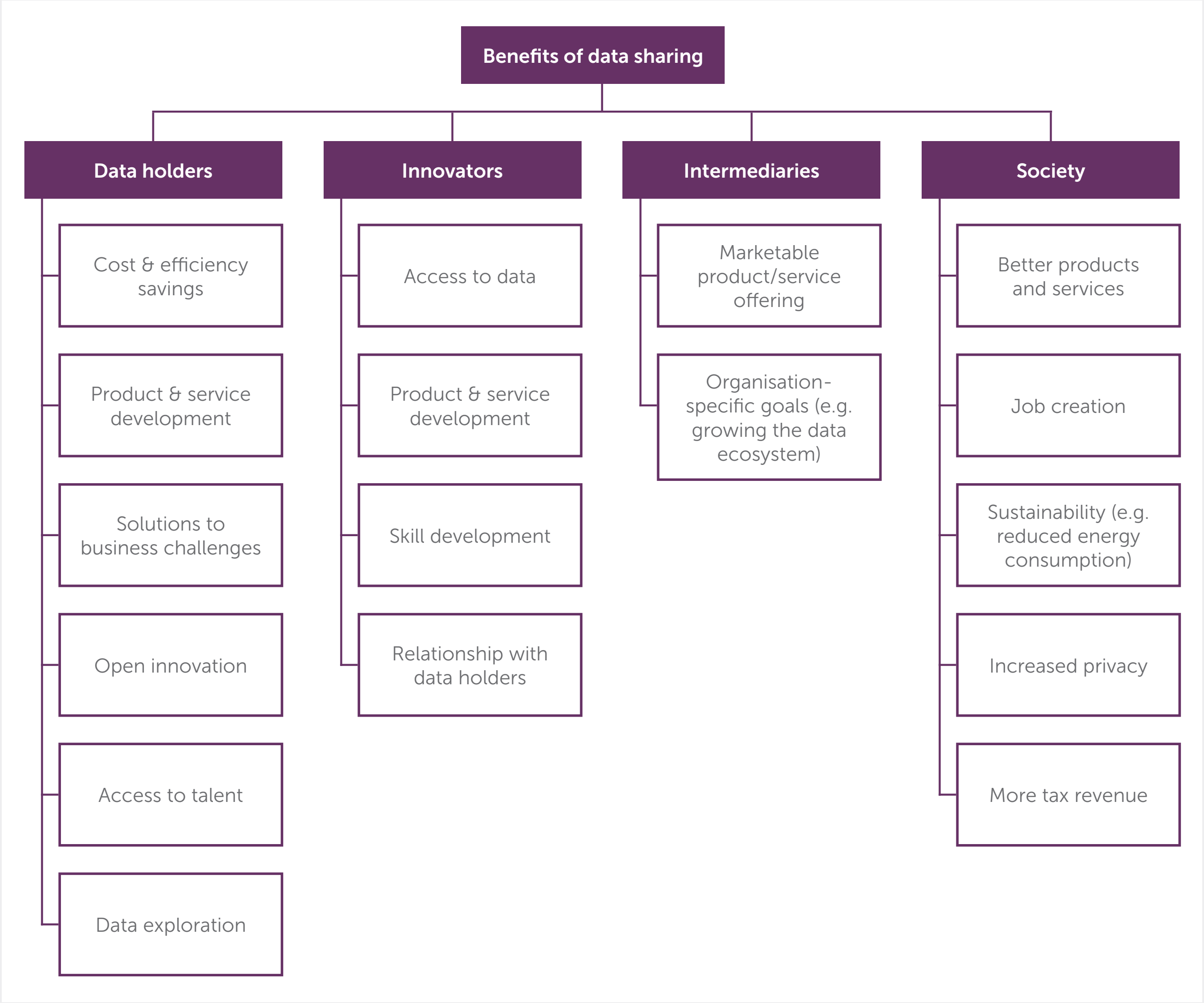
In the data marketplace, data purchasers (users) are more mature than data suppliers, often because they are organisations that have always needed more data - such as journalists, or companies seeking new customers. Key drivers for suppliers to share data are not only the creation of new revenue streams, but often also corporate social responsibility.

[9] Curry, 2016
[10] Grossman, 2016
[11] https://lo3energy.com
[12] Noveck, 2015
[13] Carnelley et al, 2016; Schomm et al, 2013
[14] O'Hara, 2019
[15] e.g. https://hazy.com/, https://www.truata.com/
[16] Hall & Pesenti, 2017
[17] Hardinges, 2018
[18] https://sidewalklabs.com

The benefits of data sharing

Sharing data has a variety of benefits in different areas, and different stakeholders benefit from data sharing in different ways.

The most obvious benefits of data sharing are economic - after all, it is often companies that hold and share data, and data users invest their time and efforts into developing *marketable* products and services. Working with data promises a variety of new products and services, jobs, business intelligence, or efficiency savings. Data sharing can also unlock benefits for the environment and to society at large, or lead to increased tax revenue.



Benefits of data sharing

Data holders
- Cost & efficiency savings
- Product & service development
- Solutions to business challenges
- Open innovation
- Access to talent
- Data exploration

Innovators
- Access to data
- Product & service development
- Skill development
- Relationship with data holders

Intermediaries
- Marketable product/service offering
- Organisation-specific goals (e.g. growing the data ecosystem)

Society
- Better products and services
- Job creation
- Sustainability (e.g. reduced energy consumption)
- Increased privacy
- More tax revenue

We will now define and discuss each of these roles - data holders, data users, intermediaries and society - in turn.

> *'Your shared data might hold the key to creating new products and services that support sustainability in transport, energy, or the environment'*

For data holders

Data holders are organisations that supply the data in a data sharing relationship. They hold - or have control over - data, and they may or may not 'own' it; for example, the data subjects might remain owners of the data an organisation holds about them, but the organisation has a right to use this data. A data holder might be a corporation or business, a department within an organisation, an NGO, a research consortium, a public entity, or any other organisation that holds data from any source.

Data holders may want to share their data for several reasons: to solve a business problem that they lack the skills to deal with in-house, to gain a competitive advantage by improving their data quality or products, or to explore what can be done with their data. Sharing the data with a data user - or possibly just with another department within the organisation - can provide new, creative ideas of what to do with or how to process the data. For data holders, the main benefit of sharing their data is to gain efficiency savings, develop new or improve existing products, create new or better services, solve existing or future business problems, or understand what is in their data. They may not have the expertise to develop these solutions internally, or it may not be economically sensible for them to work on the data themselves. By sharing data they can also get a glimpse into a developing market, in order to remain competitive.

Additional benefits of data sharing for data holders:

- improved internal data structure
- increased legal compliance
- skill development

Our work with data holders has shown that their motivation to share data can be categorised in two ways: They may be seeking a solution to a specific problem, such as improving customer recommendations, or making a process more efficient; or they may be after exploration, with the goal to find out what can be done with their data in a specific area, such as developing new value propositions or products from customer data.

Case study: Exploring data with Greiner Packaging International GmbH (GPI)

GPI was a data holder in the second call of Data Pitch. The company has a post dedicated - and with decision-making authority - to advance data driven innovation. This made it easy for them to join the programme and make a large variety of data available. Their challenge was to use the sensor data from three of their manufacturing plants to develop solutions that enhance the business in terms of manufacturing, logistics, supply chains, or even sales. This breadth of opportunity, and their capacity to fully commit to the innovation process, resulted in their partnering with five data users that are now working on advancing different areas of their business.

For data users

Data users are organisations that use data that is shared by a data holder to develop new insights, products or services. In Data Pitch, these users were typically innovative start-ups or small to medium enterprises, but they could also be another organisation, a different department in the same organisation, a university, students, individuals, or activist groups. Their main benefit in data sharing is access to data which they or their competitors would otherwise not have, which allows them to generate new insights, develop new or improve existing products or services, and establish themselves in the market. In other cases, access to vast data sets allows the data users to increase their deep / machine learning and artificial intelligence capabilities.

Additional benefits of data sharing for data users:

- business relationship with data holder (as clients, investors or other partnership)
- insight into new markets

Data users may also be data holders, and the sharing relationship in that case may be reciprocal. Organisations could swap their data, pool it for a mutual benefit, or the data user could supplement data that is shared with them with their own data. The latter was the most commonly observed situation in Data Pitch: An innovating user would combine the data provided by the data holders with their own proprietary data, to produce a solution that is mutually beneficial to both organisations.

Case study: Combining data with IPlytics and SpazioDati

IPlytics is a German start-up specialising in business intelligence. They joined the Data Pitch programme in 2018, responding to the challenge by SpazioDati, who were looking to enhance their existing business intelligence knowledge graph. IPlytics' proposed solution was to supplement SpazioDati's extensive data on various business sectors in Italy with their own data, which in turn amalgamates different sources of public data, such as patents and research publications. Their platform can be used to identify and act upon future technology trends. During the course of their participation in Data Pitch, both organisations gained a significantly improved database for their respective platforms.

> *"What excites us most is the possibility of enriching our data with the data provider's external data, as well as the possibility of having a real impact by adding value to their data in turn."*
>
> Rosann Brandt, IPlytics COO

For intermediaries

Intermediaries are not a necessary part of data sharing, but they play a role in many data sharing relationships. There are many possibly forms and roles of intermediaries, but as a general rule, they engage in-between data holders and users, and enable or help to scale the data sharing process in various ways, which we outline below.

Intermediaries will want to achieve their own specific goals, which can be defined by their members, funders, shareholders, or other decision-makers. Often this goal will be revenue: services to enable data sharing relationships are a marketable product in themselves. In other cases, such as Data Pitch, the European Commission set a goal to grow the EU data economy, to increase tax revenue, create jobs, and make the EU more competitive in data markets. Other organisations, such as NGOs or universities, might want to improve data protection, or increase the use of research data.[19]

There may be downsides to having an intermediary involved in a data sharing relationship. For example more due diligence might be required, if either public bodies or funding are involved. Such intermediaries may be held to a higher standard of accountability, increasing the required resources for checks and balances. Similarly, when an intermediary increases the efficiency in data sharing relationships, this will happen through standardisation. Consequently, there could be less freedom between the other stakeholders for the terms under which they can share data. That could be seen as a disadvantage - not using the intermediary allows the parties to draw up a contractual relationship that may fit their needs more precisely. However, this also means that they would *have to* do all of the work involved themselves, making them less able to scale their engagement.

Case Study: Enabling data sharing at the Alan Turing Institute - Data Study Groups

The Alan Turing Institute arranges week-long ‘collaborative hackathons’, which bring together talented researchers from a variety of backgrounds, such as data science or artificial intelligence, with industry problem owners. Organisations define a challenge, provide a dataset, and pay a fee to engage. The institute recruits PhD researchers from a variety of domain and data science backgrounds to work on the challenge for one week.

The data holders get to quickly prototype possible solutions to their challenges. The researchers get an opportunity to put knowledge into practice and go beyond individual fields of research to solve real world problems. As the intermediary, the institute gains industry collaborations, with the ideas generated acting as seeds that can kick-start larger collaborative research projects.

[19] Lopez de Vallejo et al., 2019

Types and roles of intermediaries

There are a variety of types of intermediaries. For the purposes of this toolkit, we class as intermediaries any third party organisation or platform that facilitates the sharing of data between one organisation and another.[20] Dependent on how much of the process they facilitate, these could range from individuals to institutions to associations. Which form they should take is currently being explored, through projects such as the data trust pilots by the Open Data Institute[21] or data collaboratives by GovLab.[22] For example, a data trust could be an institution that pools data from individuals and then negotiates terms for the use of this data on their behalf; it could be an innovation programme such as Data Pitch, which distributes development funds from the European Commission; or an institution that governs the sharing of data with a statutory oversight. It could be a framework of rules that enables data sharing, a legal construct, an organisation, or a data store.[23]

The International Data Spaces ecosystem aims to be a de facto market standard for the trade and exchange of all kinds of data assets. It facilitates the finding and authentication of appropriate transfer partners and also the legal and commercial governance of transactions.[24] The Big Data Value Association iSpaces label identifies platforms that facilitate the sharing of closed data for various innovation purposes.[25] They include the Big Data Centre of Excellence in Barcelona and the Smart Data Innovation Lab in Germany.

Different concepts of intermediaries envisage them as performing a variety of roles and consequently the benefits they provide, and the associated costs, will differ. There are a number of roles they could fill:

- **Enabling scale and capacity:** Intermediaries can help to scale data sharing relationships. Rather than developing 1:1 relationships between every data holder and user, an intermediary could bundle data holders, data subjects, or even the actual data, and make it accessible to data users at specified terms. This may entail managing the physical access to the data, or applying a framework of rules and obligations under which access to data is granted.
- **Reducing complexity:** Data sharing can be a demanding process, and at least initially be expensive. Setting up a sharing relationship takes a lot of time from a variety of internal stakeholders and experts, particularly in large organisations with complex hierarchies and decision-making structures. Intermediaries can design and apply institutional processes and regulations, and conduct due diligence checks to comply with legal requirements such as GDPR.
- **Matchmaking:** Intermediaries could identify suitable matches between data holders and users, depending on the type of data they offer or seek, and facilitate either or both of these relationships. This is what data marketplaces already do, and Data Pitch did to some degree.
- **Providing infrastructure:** Intermediaries may provide the necessary infrastructure for data sharing, although in our experience this can be treated as a commodity. There are many solutions available on the market through which data can be shared, and most data holders and users shared data through platforms or servers that they already had access to.
- **Creating trust:** Having a third party involved between a data holder and a data user can make negotiations easier. The intermediary could act as an arbiter, ensuring that both sides get sufficient benefit from the relationship, or conduct the decision making process that determines eligibility to engage in data sharing, for example by assessing data user pitches.
- **Supporting:** When an intermediaries’ main task is to grow a specific market, or address a specified set of challenges, they might focus on supporting data holders and/or user, for example by supplying templates, conducting necessary checks, or even supplying funds. Such an intermediary would likely be funded by a public body or investor, or any other entity that has an interest in growing a specific market area. Intermediaries may thus help data holders or users to acquire the skills required to engage in data sharing; Data Pitch has demonstrated this in the context of GDPR-compliance.
- **Developing best practice:** Due to their central role in data sharing relationships, intermediaries can specialise, and generate and apply best practice, making data sharing both easier and more cost-effective. This knowledge can be put to further use, for example in advising policy.

[20] However, we should note that some organisations thus classified see themselves as enablers of direct sharing, rather than as directly intermediating the relationship.
[21] https://theodi.org/project/data-trusts/
[22] https://datacollaboratives.org/
[23] Hardinges, 2018
[24] https://www.internationaldataspaces.org/
[25] http://www.bdva.eu/?q=node/790

For society

While society does not have a specific role in data sharing, it is a vital participant. There are a number of benefits to society at large that can result from data sharing. First and foremost, if innovation through data sharing improves products and services, while this will typically be motivated economically, the public also benefit from having those new or better products and services available. For example, customer service experiences are improved through chat bots or recommendations. Another area of innovation is health, where diagnoses or provision of care can be improved, and health services made more efficient and customer focussed. Data sharing can also contribute to a safer, cleaner environment, and even help tackle climate change. Many of the social benefits of data sharing double as environmental. For example, data sharing can help to improve supply chains, which in turn reduces the unnecessary transport of goods. When data is shared to develop systems that reduce emissions or energy consumption in buildings, this in turn has a positive impact on air quality and public health.

For public sector data holders, data sharing can help to achieve goals of public interest, such as more secure roads. Nine out of Data Pitch's 47 data users aimed for environmental effects, for example by improving traffic flows or maintenance works, both of which could contribute to both better services, a safer urban environment, as well as reduced emissions. Safety was the focus of a number of other projects, which aimed to give citizens better control of their data, improve their privacy, and organisations' compliance with the GDPR. Enhanced insight through data can also be used to make more socially-responsible and environmentally-friendly decisions within organisations, and lead to better policy decisions among state actors.

In a wider sense, society or the data community benefits from data sharing, as new jobs are created. If new algorithms or other AI insights are published openly, the ecosystem can benefit further as learning is accelerated. This will increase awareness, as well as the quality of data and data processing. Case studies of successful data sharing will also make other organisations more likely to engage in future data sharing activities, increase the availability of data for everyone, grow the data ecosystem, and turn data sharing into a more common practice.

In the next section of the toolkit we demonstrate how data holders, users and intermediaries interact to create economic value and develop the data ecosystem in Europe, through a case study of the Data Pitch data innovation programme.

Case study: Met Office addressing air quality with GoSweat and Hop Ubiquitous

Data users GoSweat and Hop Ubiquitous are working with MET data provided through Data Pitch, to improve the use of pollen and air pollution data, and ultimately impact public health. Hop Ubiquitous is building a decision support system to help public servants and citizens make more environmentally aware decisions. GoSweat's application allows end users with hay fever to plan their exercise around pollen forecasts.

"We believe in sharing our data and enabling others to use it. We see our involvement with Data Pitch as a key to making data more available and usable. [The ideas developed through Data Pitch should support] UK citizens, by making life easier, protecting them, helping them prosper or improving well-being."

Richard Carne, Chief Digital Officer, Met Office

Case study: Deutsche Bahn and Ubiwhere enhancing transport flows for economic and environmental benefits

Ubiwhere uses data provided by Arriva (the UK arm of Deutsche Bahn) to improve their mobility solution. Combining historical booking data, with data about external factors, such as weather, points of interest, or seasons, they provide business intelligence that will allow their customers to identify and proactively resolve inefficiencies.

"We want to improve the punctuality of our bus services based on outside influences such as traffic flows, weather, events and unforeseen incidents and how they impact schedules, and to see what we could do to be not only reactive, but also proactive – so for example if a car breaks down on one of our routes we can find out in advance and address it with diversions for following buses. Ubiwhere used external data points to see how they affected journey planning. This included traffic light networks, crowdsourced data for traffic flows and weather patterns."

Stuart Walker, Senior Product Manager, DB/Arriva

ABOUT DATA PITCH

Data Pitch was a Horizon 2020 Big Data Value Association Public Private Partnership open innovation programme, bringing together data holders - corporate and public-sector organisations that have data - with data users - startups and SMEs that use data.

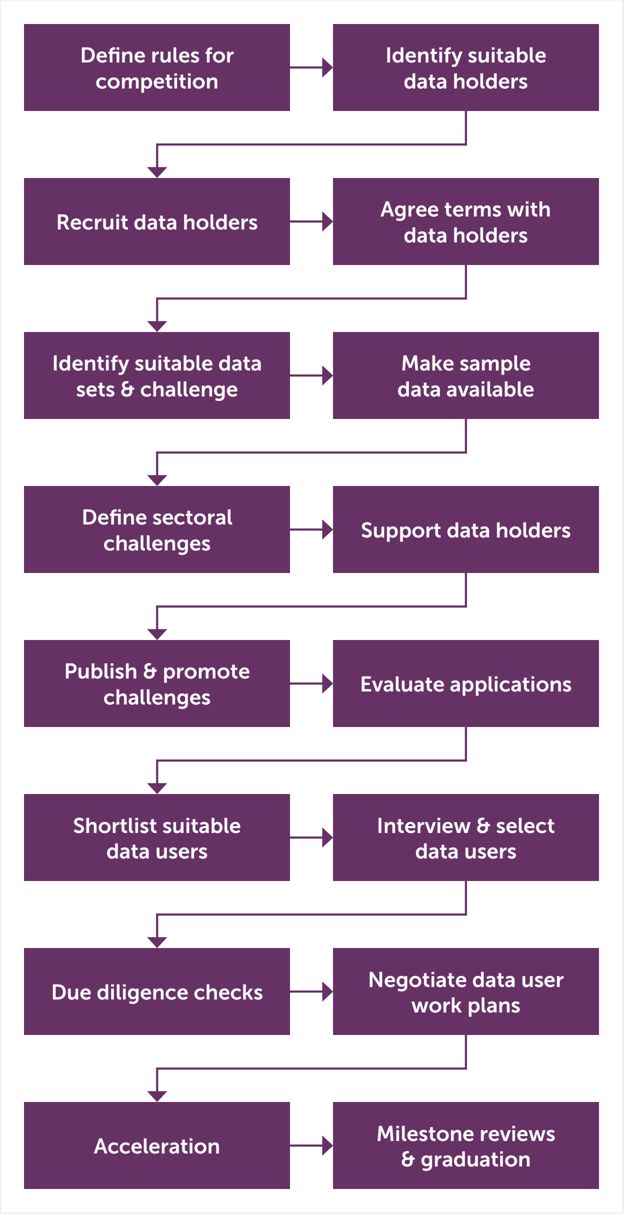
It was funded by the European Commission, in order to nurture the data ecosystem in Europe. At the heart of Data Pitch lay the concept of *data sharing:* organisations made their otherwise closed data available for the development of new products or services.

The programme provided a supportive environment in which both data holders and users could learn and experiment. It facilitated data sharing through a competitive call mechanism, using challenges in sectors that were of particular importance for the EU, such as finance, transport, and energy. Two types of challenges were defined, in collaboration with industry experts and data holders:

- **Data holder challenges:** Data Pitch sought out organisations who were willing to formulate a challenge based on business problems or data they wanted to explore for potential use. They agreed on a purpose and suitable terms under which they would share data, and supplied sample data through a dedicated platform. Since these challenges were tailored to the business needs of the data providers, data users had to use the data that these providers were willing to share, and responded to the challenge with potential solutions they could develop based on this data.
- **Sectoral challenges:** Data Pitch defined broad challenges in a wider range of sectors, based on desk research and industry expertise, and focused on the BDVA key verticals, but did not contract with data holders to provide data. Data users submitted proposals that addressed those challenges, confirming that they would complete them using data that is provided by someone else, that is not open, and to which they had access, for example via a license or any other type of agreement. As Data Pitch did not need to manage these relationships directly, it allowed the programme to scale.

Data holders benefited from exploring their data, introductions to the machine learning ecosystem, open innovation, experience and guidance in the process of data sharing, and a time bound licence to the results which could address a company problem.

Data users benefited from access to data, 100,000€ of funding for their solution, and a 6-month accelerator with dedicated support. Though the financial support was certainly a major incentive for them to join the program, the access to data was valued even higher.


Define rules for competition
Identify suitable data holders
Recruit data holders
Agree terms with data holders
Identify suitable data sets & challenge
Make sample data available
Define sectoral challenges
Support data holders
Publish & promote challenges
Evaluate applications
Shortlist suitable data users
Interview & select data users
Due diligence checks
Negotiate data user work plans
Acceleration
Milestone reviews & graduation

How it worked: Defining challenges

The central data sharing mechanism around which Data Pitch was built was the open innovation challenge answered by a competitive call. Consequently, the first task was to define the rules for that competition - who could set a challenge? What kind of challenges would they be? Who would be eligible to answer the challenge? How would they be selected? What timescales would this take?

Challenges were defined by public sector and private sector organisations from across Europe, who shared their previously closed data for the purpose of enabling skilled, innovative companies to provide solutions to those challenges. In Data Pitch terminology, the companies who shared the data were known as *data providers,* and the companies that aimed to use that data to solve challenges were known as *solution providers.*

Data providers received support and service in the process of data sharing, and a time bound licence to the results the solution providers produced. Data Pitch worked closely with them to help them understand the benefits and risks associated with data sharing. This was partly due to the fact that many of the data holders did not have much previous knowledge of sharing data, and therefore needed to establish its value across the organisation before things could move ahead. Typically starting from one contact in a single department, the idea of Data Pitch, as well as the terms of the programme, had to be communicated through multiple departments in order to secure agreement from a business strategy, technical, and legal perspective.

The discussions involved in both defining the challenge and selecting appropriate data to address it, required significant resources. This was because the more complex the organisation was, the more different departments needed to be involved. For example, the legal department would assess whether and which data could be shared legally; a technical or data department would assess which data was available or could be shared technically, and what processing would be required before it could be shared; and a business or strategy decision-maker would need to sign off the engagement in the process in general, which would be dependent on the potential benefit gained from this engagement, the required resources, and the challenge it addressed. Assembling a team and having established communication across all of these departments proved to be crucial for success. This process became easier the more focused the initial contact was, the more decision-making power the internal stakeholders had, and the more experience organisations had with either innovation with data, or open innovation.

Onboarding a data holder and supporting them through this whole process took a dedicated team within Data Pitch up to one year. It was not always successful: Some organisations, where discussions had been active and fruitful for months, had to withdraw from the process. This was for a variety of reasons, including reorganisations, a focus on 'business as usual', or a risk averse corporate legal culture.

As tempting as it was for a data holder to simply release data sets and set a challenge such as 'Tell us what the value is in the data set' or alternatively, to define a challenge and then later decide which data sets would be the most useful for answering it, it was crucial for compliance with data protection regulations that the two informed each other.[26]

As the data holders had to agree to share their data with organisations who would address their challenges before those organisations had been identified, Data Pitch developed a bilateral, asynchronous contract, or in fact, two contracts. The terms of the data provider contract mirrored those of the contract with the solution provider, but were executed before the launch of the open call. The second contract was executed after the selection process was closed.

During the process of recruiting data providers, two things became apparent: many organisations did not have the resources to work with more than one or two startups at a time; at the same time, Data Pitch would have the capacity to support more data sharing innovation. To enable this, we introduced 'sectoral challenges', in which Data Pitch oversaw the definition of important industry challenges and invited entrepreneurs to solve them, using shared data they sourced themselves. This enabled Data Pitch to support a large number of innovation projects covering a much wider range of application contexts.

Recommendation:

Data sharing is an area that many parts of an organisation may feel they have a considerable stake in. Representatives of the technical, legal and business sides should all be involved in the discussion as early as possible. This will allow them to take an equal part in the conversation and framing of the data sharing relationship. Prioritising one aspect over the other may lead to increased risk aversion from the other parts of the organisation.

[26] More about data protection compliance can be found in Stalla-Bourdillon et al. (2019)

A Data Pitch challenge and associated dataset

Challenge identifier: DPC3-2018

Proposed by Greiner Packaging International GmbH (GPI). GPI manufactures and markets plastic packaging solutions for food and nonfood industries.

Background

Sensor data and other Internet of Things (IoT) technologies used in industrial processes (known as the Industrial Internet of Things (IIoT)) are providing manufacturers with increased opportunities to optimise their operations and business processes, as well as engaging with their customers.

The smart factory not only represents a step forward from traditional automation, but also rests on a fully connected and flexible system—one that can use a constant stream of data to learn and adapt to new demands.

Manufacturers who are able to access such insights are able to optimise business and manufacturing processes better than ever before. The market for big data manufacturing software is estimated to be one of the largest opportunities in any industrial category.

Description

GPI has recently invested in extensive sensors across 3 manufacturing plants. We are now looking for ways to utilise this data, along with our existing data, to best support and enhance our business. We are looking to develop applications which span across traditional boundaries of manufacturing, logistics and supply chain (perhaps even sales), providing data-informed solutions to better coordinate these processes, make them more efficient, and platforms supporting these new processes.

We are particularly interested in solutions that:

- Define new relationships between data to provide new insights and understanding;
- Discover new business opportunities and improve production efficiency;
- Help ensure the integrity of interactions within supply chains;
- Integrate data of different modalities (sensors, acoustic data, historical records, thermal maps) to produce useful products and services;
- Use data to predict maintenance needs and create more efficient servicing and repair services; and
- Predict and help optimise consumption and stock level, including in multi-country operational scenarios
- Capture and interpret data to produce answers to commonly asked questions and reduce human intervention;
- Are able to integrate poor, inconsistent or fuzzy data or information and provide interfaces that communicate key findings and effectively engage users.

Data

Examples of data include but are not limited to:

- Production orders
- Logistics (order process)
- Sensor data in production (machine, energy consumption, cooling water)
- Environmental data (shop floor temperature/humidity)
- Quality data (product properties, scrap rate)
- Failure cases of machines
- Maintenance and usage history

More detailed information about the data can be found in our data catalogue.

Expected outcomes

Examples of outcomes include but are not limited to:

- Prediction algorithms that help decrease total stock holdings and lost sales
- Supply chain optimisation algorithms
- Algorithms and applications that integrate different sources of data in interesting and novel ways
- Insights across business functions such as marketing, operations, product and development, sales, etc.
- Repeatable analytic processes that accelerate the adoption of analytics
- Ability to gain a better understanding of what data is currently not collected
- Ability to develop benchmarks that over time contribute to the optimisation of future decisions
- Waste and lead to better carbon footprints
- Reporting, analytics and visualisation tools that help to:
  - Absorb information in new and more constructive ways
  - Visualise relationships and patterns between operational and business activities
  - Manipulate and interact directly with data
  - Allow other stakeholders to engage with the data

Expected impacts

- Ability to make better informed decisions e.g. strategies, recommendations
- Ability to discover hidden insights e.g. anomalies forensics, patterns, trends
- Facility for automating business processes e.g. complex events, translation
- Performance payoffs
- New business processes
- Improved decision making

How it worked: Selecting solution providers

Having defined the challenges and signed contracts with data providers, the call was launched. European small to medium enterprises (SMEs) applied through F6S,[27] a platform designed to connect tech founders and startup programmes. Eligible proposals were assessed by two reviewers who shortlisted candidates for interview. Data Pitch evaluated proposals, shortlisted and selected the most suitable data users through an assessment of their idea, team and budget, and the anticipated impact of the proposed solution.

SMEs who passed the interview stage then entered a negotiation phase to become solution providers, where a team of advisors worked with them to develop detailed work plans. At the end of this phase the solution providers gained access to the shared data, received 100,000€ of funding, and commenced working on their challenge for a period of 6 months. At the end of the 6 months, the data reverted to the data providers, while the intellectual property of the innovation remained with the solution providers, who granted a one year, non-commercial, non-exclusive licence on the results of the innovation to the data provider.

While not all of the solution providers' ambitious goals were achieved, this did not mean a lack of success. To the contrary, often the results were more useful to data providers than originally intended.

Where simpler solutions were delivered, these were successful not despite but because of their simplicity, and simultaneously more marketable and easier to maintain for the solution providers. Along the way, data providers also gained more insight, knowledge and process improvement. Examples of this included: awareness that knowledge about GDPR was not simply restricted to the legal team and that the technical staff should also receive training; the automation of a difficult data merge (with no common identifiers) that previously had to be completed manually by skilled staff; identification of the documentation necessary to use the data effectively, and codification of the staff experience that was required for the data to be used.

Key resources:

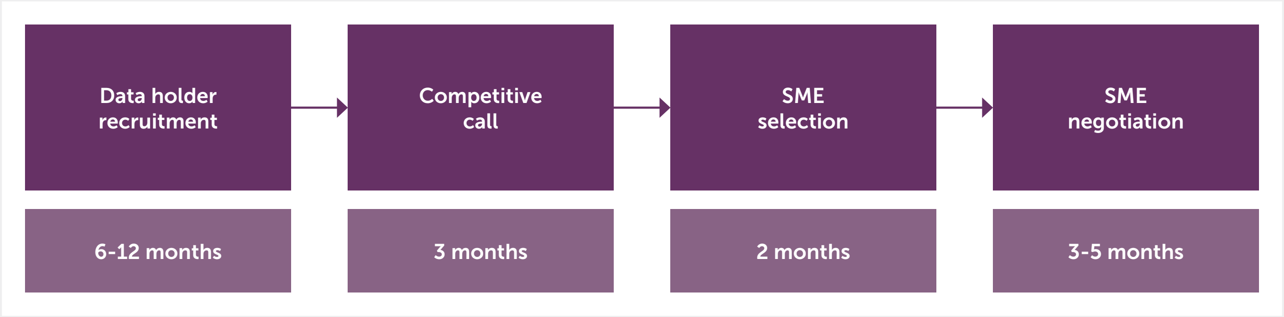
Data Pitch reports: All the processes in Data Pitch are documented in a series of reports, which are publicly available. This includes how challenges were defined, how data was managed and the development of the data catalogue.

All resources can be accessed at *https://datapitch.eu/deliverables/*

[27] https://www.f6s.com/datapitchaccelerator

Cost of data sharing

The European Commission funded the Data Pitch consortium with 2.5 million Euros. Roughly half of this budget was spent on finding and onboarding data holders, running the competitive call, and the selection and onboarding of SMEs.



The cost accrued for running the Data Pitch programme must be considered together with the intended scale of the programme. Data Pitch sought to identify and on-board up to 50 data holders across 15 sectors, and select at least 50 SMEs to work with them, all in a period of three years.

Discussions were initiated with three data holders for each data holder that was successfully on-boarded. For the 13 data holders that joined the programme, a team of business, technology and legal experts in both organisations worked through the challenge that should be addressed, the data that could be used, and the legal framework, to ensure that all parties' requirements were met.

For each of the 47 SMEs that joined the programme, five times that many applications were reviewed. Every successful SME went through a negotiation period to confirm their work plan and budget, which was then assessed at three milestone reviews over the course of the acceleration phase.

Some parts of Data Pitch scaled well, such as the governance process, which was defined at the beginning and then re-used - contract templates, once developed, needed hardly any changes; the challenges, once defined, needed only minor tweaks and some ongoing support for applicants.

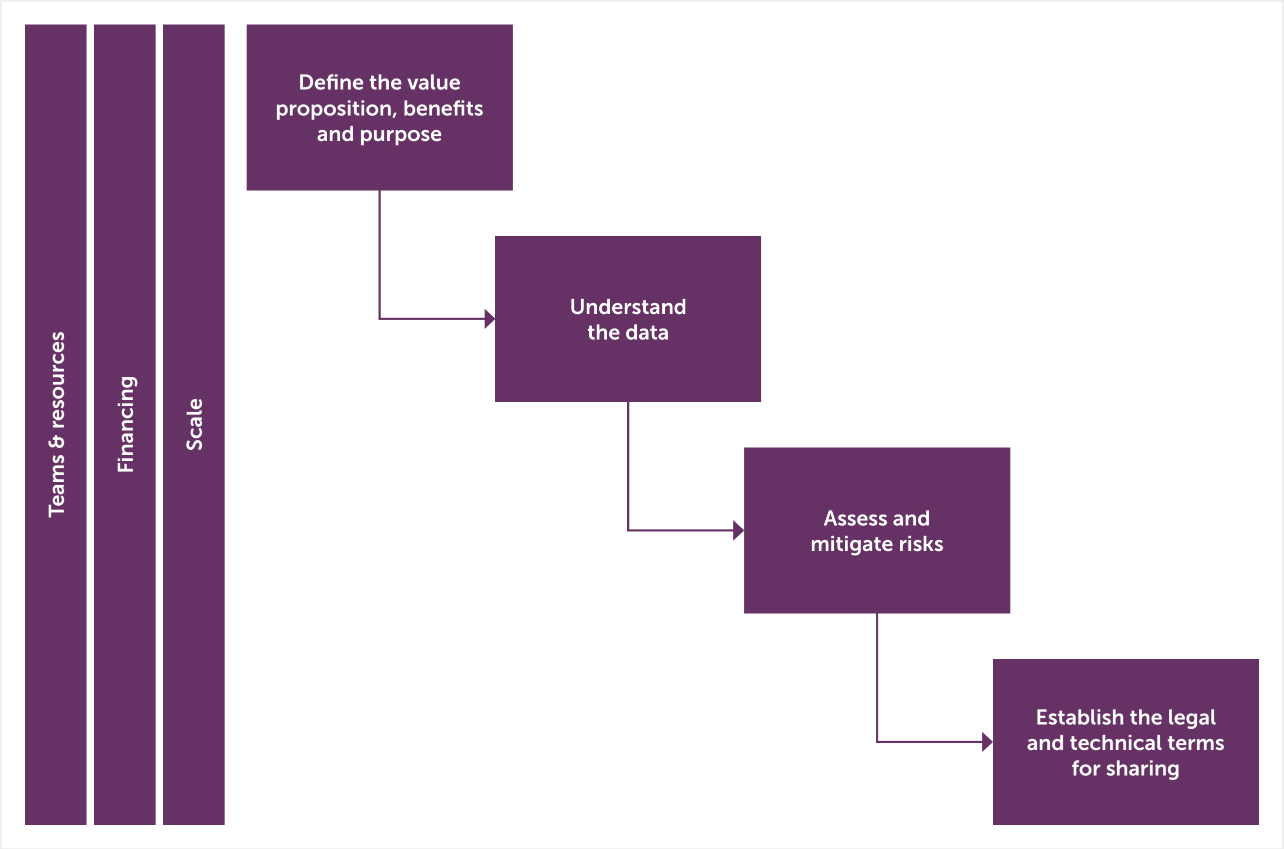
Other parts, such as the identification and preparation of data with data holders, pushed the limits of capacity, as they needed to be completed individually for each data holder, sometimes more than once.

The workload for Data Pitch as the intermediary increased with both the number of data holders and users engaged in the programme, especially as Data Pitch was part of all data provider-solution provider relationships for those challenges. Our data providers on the other hand worked with one to five SMEs simultaneously, and their resource commitment increased with both the number of datasets they shared, and the startups/SMEs they engaged with.

'Sharing your data outside your organisation is a proven way to optimise processes and systems inside your organisation'

HOW TO SHARE DATA

The next section of the toolkit lays out the process of sharing data, and presents the decisions and considerations that have to be made at each step.


Teams & resources
Financing
Scale
Define the value proposition, benefits and purpose
Understand the data
Assess and mitigate risks
Establish the legal and technical terms for sharing

'If you want to discover a whole new world of innovation, get your data in order and develop technical and legal knowledge in your organisation, data sharing should definitely be on your horizon'

Define the value proposition, benefits and purpose

"People think the actual flow of data can be a barrier. It's not really. As soon as people decide ... the problem is getting to the decision of sharing the data."
(Head of Startup and Innovation Programmes, Open Data Institute)

Before any data sharing can commence, the value proposition of the data sharing relationship - exactly why data is being shared and who with - requires defining. We discussed some of the benefits of data sharing on page 7. The first step toward successful data sharing is to clearly define its purpose. There should be a document outlining how all parties will be engaged. Questions that this should address are:

1. What is the benefit of getting involved for each of the intended organisations or individuals?
2. How is it ensured that these benefits are realised?
3. What does each of the stakeholders have to commit to realise them?
4. Will any external stakeholders receive benefit/need to be involved?
5. What terms and licence will be needed to ensure these benefits accrue to each party?

This does not mean that the exact desired result has to be known upfront - this will rarely be possible - but a general understanding of the intention of the parties involved is necessary. While it might be only one party that defines this initially, it is important to agree what success looks like; this will flow through to other parts of the process, such as the definition of KPIs or other measures of success.

Aside from benefits to the involved organisations, it is also helpful to consider what the *individuals in charge within these organisations* may want to get out of the process. For example, what motivates a technology manager in a large corporation to engage in data sharing may differ from what motivates the owner of a small enterprise. Data sharing is more likely to be successful if the decision-makers across the involved organisations agree what success looks like.

There are different ways how data holders and users can find one another. For example, a data holder could publish a challenge or a call for tender for data users to apply to, or directly approach a data user of their choice.

Key resources:

A framework for data sharing for open innovation (Walker et al., 2019): Outlines a framework that identifies the conditions which enable value to be created through a data sharing format.

https://drive.google.com/file/d/1PsJo4v4NEqp5ANe1zwfjDEzERM5qh1wr/view (p. 112)

Data collaboratives (Verhust et al., n.d.): Comprehensive overview of data sharing institutions for public good, spearheaded by the US-based GovLab.

https://datacollaboratives.org/static/files/data-collaboratives-intro.pdf

Data trusts: lessons from three pilots (Hardinges et al., 2019): Report by the Open Data Institute on their work to establish what Data Trusts are and how they can function.

https://theodi.org/?post_type=article&p=7888

Creating the energy data commons (Webb, 2018): Overview of a data commons model in the energy sector.

https://lo3energy.com/creating-the-energy-data-commons/

Understand the data

"The issue when people decide to investigate data is often that data in a spreadsheet or an unstructured mass generally looks fine. It is once you start to try and utilise it in some way, or apply some kind of interrogation to it, that you realise what the problems with the data are."
(Data Provider Liaison Lead, Data Pitch)

The data and the issue that will be addressed using the data need to be well aligned. This will require a combination of expertise from different areas, including:

- understanding whether an intended outcome can be achieved through the use of data;
- knowledge of what data will be needed;
- knowledge that this data exists, and where;
- knowledge of what the data includes - the metadata to the data, such as its level of granularity, to assess its usefulness;
- access to both metadata and actual data; and
- authority to share both metadata and actual data.

This may require the production of an information asset register or inventory, even if only a subset of the organisation's data is used.

In order to explore the data, especially if there is limited awareness of the potential uses, it may be useful to provide sample data. Access to sample data allows potential data users to judge whether or how their proposed solution will be achievable. Alternatively, a synthesised subset of the data may be made available, to allow exploration of the data without risk of breaches.

What is synthetic data?

Synthetic data is a disguised version of a subset of the original data, which no longer holds any real data. In all other ways it appears to be the real data in terms of metadata, noise and other features of the original dataset, the most relevant being the statistical distribution. Synthetic datasets can be used to train algorithms that can then be tested on the real dataset, without revealing real data.

Data can be fully or partially synthetic. Fully synthetic data does not contain any original data; re-identification of any single unit is almost impossible, while all variables are still fully available. In partially synthetic data, only data that is sensitive is replaced. This leads to decreased model dependence, but does mean that some disclosure is possible owing to the true values that remain within the dataset.

Key resources:

Valuing information as an asset (Higson & Waltho, 2009): This white paper aims to support senior executives and policy makers with the transformation of information culture and practices within their organisations.

http://faculty.london.edu/chigson/research/InformationAsset.pdf

Data inventory guide (Johns Hopkins University, Center for Government Excellence): This is a practical guide to understanding what a data inventory is and how to build one, explaining the concepts and providing practical guidance and references.

https://labs.centerforgov.org/data-governance/data-inventory/

Designing data governance (Khatri & Brown, 2010): This paper offers a framework for data governance issues to help practitioners design data governance structures effectively.

https://doi.org/10.1145/1629175.1629210

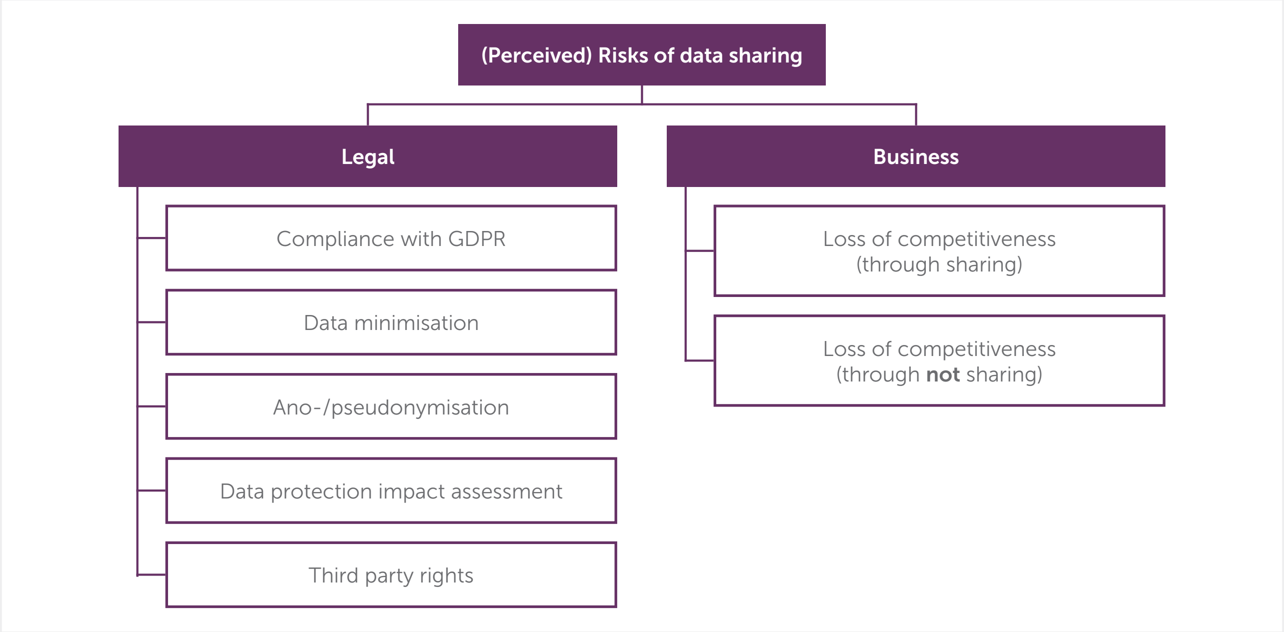
Anonymisation and open data: An introduction to managing the risk of re-identification (Thereaux et al., 2019): This report describes the current state of the art of data anonymisation and provides guidance for risk management.

https://theodi.org/article/anonymisation-and-synthetic-data-towards-trustworthy-data/

Assess and mitigate risks in the use of data

"Most of the reasons for not sharing data were not about 'We don't believe that we can do something with our data', but about 'We are not prepared', 'We do not want to take the risk', or 'The legal office will not sign.'"
(Business & technology advisor, Data Pitch)

An assessment of the risks of engaging in a data sharing process is required for all stakeholders in a data sharing relationship, but most important for data holders, as they are accountable for the use of the data that they share. There are several reasons why organisations do *not* share data. Two of the most common are: the legal compliance and associated required processes of sharing data; and the competitive advantage that could be lost or gained in the market through data sharing.



(Perceived) Risks of data sharing

Legal

- Compliance with GDPR
- Data minimisation
- Ano-/pseudonymisation
- Data protection impact assessment
- Third party rights

Business

- Loss of competitiveness (through sharing)
- Loss of competitiveness (through **not** sharing)

Legal risks

No data can or should be shared without ensuring legal compliance, and conducting a legal risk assessment. Initially, this requires organisations wanting to share data to be familiar with the legal framework that applies to them. For data shared in Europe, this will be the General Data Protection Regulation (GDPR). With its introduction, the potential repercussions for breaches of privacy and data protection can amount to up to 20 million Euros, or 4% of global annual turnover.

However, done well, data sharing need not be a risky activity. To the contrary, the potential risk of sharing data should be balanced against the organisational risk of *not* realising the value of the data. The cost of risk mitigation may be comparatively small, if the potential benefits are significant efficiency savings, or a new product in a competitive market.

Data protection impact assessments (DPIAs) are mandatory for personal data processing that "is likely to result in a high risk to the rights and freedoms of natural persons" (Article 35 of the GDPR). Even when this is not the case, they are still best practice.

Data subjects are living individuals that are identified or identifiable through a specified data set i.e. personal data (as defined by Article 4(1) of the GDPR). One way in which a dataset can relate to data subjects is through its content. In some cases, it is extremely obvious that data relate to data subjects – e.g. when the data contains information about Maria Smith, such as her name, date of birth, and contact details. A dataset can also relate to data subjects through the purpose for its processing (e.g. to learn something about or evaluate an individual), or the result of its processing (i.e. it is likely to have an impact on the rights and freedoms of data subjects).

In many instances of data sharing, the data shared is or could become personal data, and the same dataset can be considered as non-personal or personal under different circumstances. Furthermore, there is a risk that anonymised data may be re-identified.

Other legal questions that need to be addressed are that of consent, and of third party rights to the data. If the data was collected from a natural person, even if they are no longer identifiable, did the data holder obtain consent at the time of collection for the data to be used for the intended purpose? Is it a purpose that the original person could reasonably believe their data might be used for? Finally, in order to be able to share a dataset, data holders must ensure that they have appropriate license over the data to do so. There should be no additional third party rights, such as others' intellectual property contained in the data. While these might seem onerous, establishing clear answers greatly facilitates successful data sharing.

Some of the questions data holders should consider in this context are:

- Is personal data being shared?
- Has the data been anonymised or pseudonymised?
- Have data minimisation principles been applied?
- Are the safeguards and measures in place adequate to control the flow of these data?
- Is the data processing considered high risk? Is a DPIA required?
- Which data subjects' rights are concerned, and what usage of their data have affected data subjects consented to?

A note on data protection

The legal and privacy toolkits developed by Data Pitch can help work through these considerations, and provide guidance on:

1. The types of data that are likely to fall within and outside the scope of the GDPR.
2. The types of data processing activities that are considered as high-risk under the GDPR.
3. The types and levels of measures that are required to control the flow of data.
4. The basics of data flow mapping as an approach to the creation of data situation models – to be used as part of anonymisation assessments.

> *'To make machine learning a success in your organisation, you have to address red tape and legal roadblocks head on.'*

Business risks

Putting aside data that sends anti-competitive commercial signals to the market, there are other potential business risks in sharing data. There may be a risk of losing a competitive advantage when data is shared. Sharing data always implies a potential loss of control over the data, as data leaves organisational boundaries. On the other hand, there is also a risk associated with not sharing data, and foregoing opportunities to develop and gain a competitive advantage in the process. This is a strategic business decision: that continuous development is necessary to stay competitive, and data sharing is a suitable and cost-effective way of doing this.

The more competitive a domain is, the higher is the perception of risk when data is allowed to leave the boundaries of the organisation. Whether data sharing affects the competitiveness of an organisation - one way or the other - will largely depend on the trustworthiness of the data users it is shared with and how that trust is structured. An assessment of the potential of data sharing, and an appropriate vetting of data users, are key to risk minimisation. In order to ensure that data users are trustworthy, data holders should conduct sufficient due diligence checks.

Another risk is a failure in the processing of the data, which might in turn be caused by insufficient quality or quantity of data. As above, the release of a sample or subset can assist with this. Making sample data available also helps to make the engagement between data holders and data users more fruitful, as questions can be raised early, and data or solutions can be refined to accommodate them. Finding out that more data is needed, or the data itself needs extensive pre-processing before work can commence, can be a significant barrier. Continuous discussions between data holder and user can help mitigate this risk.

A note on pre-processing

Pre-processing, especially on large, unstructured data sets, can consume extensive time and resources. It is important to establish who will be carrying out this work as part of the data sharing arrangement. Depending on the amount, quality and status of the source data, this may be a substantial amount of work, which should be estimated and costed upfront. In a Data Pitch survey, data users described that low data quality and required data preparation were the biggest and most underestimated challenges they experienced.

Key resources:

Legal and privacy toolkit (v1) (Stalla-Bourdillon & Knight, 2017): A guide that focuses on the critical things to consider when sharing and reusing data for defined innovative purposes under the Data Pitch programme. Including an overview of the legal and regulatory framework that applies to data sharing and data reuse, and a risk mitigation strategy for the secondary use of data.

https://datapitch.eu/privacy-toolkit-v1/

Legal and privacy toolkit (v2) (Stalla-Bourdillon & Carmichael, 2018): A guide on the basics of mapping data flows as an effective and practical approach to the creation of data situation models for anonymisation assessment and GDPR compliance.

https://datapitch.eu/privacy-toolkit-v2/

Data protection by design: Building the foundations of trustworthy data sharing (Stalla-Bourdillon et al., 2019): This paper suggests a common workflow to embed data protection by design within data sharing practice.

https://dx.doi.org/10.5281/zenodo.3079895

Anonymous data v. personal data—a false debate: an EU perspective on anonymisation, pseudonymisation and personal data (Stalla-Bourdillon & Knight, 2017): This paper discusses the benefits and challenges of data anonymisation.

https://eprints.soton.ac.uk/400388/

European Commission regulation on the free flow of non-personal data (European Commission, 2019c)

https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data

Establish the legal and technical terms of sharing

"There was a lot of discussion over issues like the jurisdiction or the length of the intellectual property rights. Occasionally you would go back and forward for weeks on end, and then realise that it was just a language barrier."
(Data Provider Liaison Lead, Data Pitch)

The risk assessment, the purpose and selection of the data feed into the terms that are set down in the contract(s) governing the data sharing relationship. Beyond general terms, such as the jurisdiction for resolution of any disputes, the contract might include terms around licenses for the data, data protection obligations, and intellectual property rights in the results.

Depending on the stakeholders involved, especially whether an intermediary is involved or not, the signatories to the agreement may be different. If the data sharing relationship is set up directly between a data holder and a data user, a simple bilateral contract will suffice. An intermediary may broker a relationship between data holders and data users so that they can sign a bilateral agreement; or they may be involved in the contractual relationship itself. While this adds complexity to the relationship, it also has benefits, such as the option to select data holders and data users at different points in time (as was the case with Data Pitch), or the intermediary acting as an arbiter, ensuring both data holder and data user gain sufficient benefit from the relationship.

Data sharing arrangements can be one:one, one:many, many:many and, conceivably, many:one. Beyond the general due diligence checks to ensure the data is shared between legitimate parties, the exact terms of this relationship need to be negotiated. The main concern for data users as well as data holders will typically be the terms under which the data that is shared (its purpose of use, length of time, etc.), and ownership of any outcomes. A data sharing agreement can set out what the shared data includes and which anonymisation or pseudonymisation it has been subjected to, how and when the data will be shared, and the obligations and liabilities of all parties, especially with regards to data protection, and what happens with the data upon completion of the contract.

Data holders should decide what terms they can accept to share data. Considerations could include:

- Who owns the intellectual property rights in the outcome of a data sharing process?
- What outcome is required in order to justify the required resources?
- What data do data users get access to, and for how long?
- How will any disputes get resolved?

This negotiation will be an iterative process, where the data, purpose of use, and implications, are revised in context; this requires expertise that may have to be brought in, especially since GDPR is not yet well understood in many organisations. Data users may need support in identifying and mitigating risks related to the use of personal data. It is in the interest of both data holders and intermediaries to ensure that this is provided, as all stakeholders may be liable in case of a data protection breach. All parties to a data sharing agreement need to be aware of the data protection implications of the shared data, to ensure they follow procedures and minimise the risk of breaches when the data is used.

As well as the legal framework for sharing, the technical framework must be addressed. How will data users access the data? Where can they work on it? What processes must be undergone at the end of the data sharing period (as defined in the legal agreement)? As shown earlier in this toolkit, there are a number of data sharing spaces emerging that offer platforms and tools for data sharing. However, successful data sharing has also taken place through more commoditised routes such as via Amazon Redshift or the Google Cloud.

Key resources:

Data legality reports: These reports reflect on the process and considerations in setting up two of the most crucial contract templates for use in Data Pitch.

https://datapitch.eu/data-legality-report/

Data sharing agreements: As part of the Data Pitch programme, asynchronous bilateral data sharing contracts were created. These are the SME Contract which sets out the terms and conditions of how the Data Pitch consortium engages with SMEs for the duration of the programme, as well as the data provider (data holder) agreement which sets out the terms and conditions for those organisations that are sharing data. All of these are available for information and inspiration:

- **For data holders:** *https://datapitch.eu/DSA-data-holders/*
- **For data users with data shared by a Data Pitch data holder:** *https://datapitch.eu/DSA-dp-data/*
- **For data users who recruited their own data holder:** *https://datapitch.eu/DSA-own-data/*

A note on teams and resources

As we have shown, relatively substantial business, technical and legal knowledge about an organisation's data is required for successful data sharing. Consequently, enabling data sharing requires empowered leadership of the process by an individual or team that has access to resources and the cooperation of other teams in the company.

While some individuals may have insight or experience that would be valuable, often this knowledge is tucked away in separate departments, and not available to the wider organisation or those who are involved in the decision about data sharing - such as strategists, lawyers, analysts or technology managers. This lack of expertise may in turn lead to misconceptions of the risks associated with data sharing.

In this section we discuss the personnel and culture that are necessary and effective for data sharing.

Data holders

Depending on how an organisation is structured, a data holder is likely to have at least three areas (most likely represented by even more people) involved, which between them cover

- the technical aspects of the process, such as identifying suitable data and implementing the results of the data sharing process;
- legal questions, including the assessment of the legal risk of the process, and supporting the signing of contracts;
- business or strategic expertise, to assess the business risks and suitability of the project to the organisational goals.

In each of these areas, it is important that either the team working on data sharing has decision authority, or the responsible senior decision-makers are sufficiently informed to sign off the process as a whole. Ensuring the buy-in of the final decision-maker early on is vital for success.

This team also needs to have the authority to overcome other barriers. A senior-level, strategic decision for data sharing is a necessity, especially with regards to potential risks. Nothing less than a strategic decision that the organisation cannot afford *not* to make data sharing work will overcome the concerns of a completely risk averse legal department.

While some organisations have dedicated innovation or data science teams which can drive the data sharing process forward, other organisations may lack either the internal culture in which new ideas such as data sharing can thrive, an awareness of the benefits of engaging in data sharing, or the decision-making structure that is needed to enable data sharing.

If the organisational culture does not encourage or even permit developing new ideas, enabling data sharing may require a change in culture or attitude. Different internal stakeholders will need to understand the benefits of data sharing, both in general and in the specific proposed context. Decision-makers will need to be (made) aware of the value that could be derived from the data. Without this awareness they are not likely to consider sharing the data; without knowing the potential value it will be impossible for them to weigh it against the potential risks, and thus make an informed decision.

The necessary awareness and knowledge will have to be built up in the organisation as part of the process, such as:

- learning about data sharing and the value proposition attached to it;
- definition of challenges or business problems that could be solved with data;
- identification of suitable datasets, and legal assessment of whether and under which terms data can be shared;
- sign-off for the engagement

This is not, and cannot be, a linear process, as all of these steps inform one-another, and therefore will have to happen somewhat simultaneously.

It takes time and effort to build the necessary knowledge and forge the internal pathways within organisations that are required for successful data sharing. How difficult this learning process is will depend on the amount of resources that are committed to it, which is in turn dependent on the buy-in that data sharing has in the organisation. All of these may have to be built up in an iterative process, and it may be that utilising the services of an intermediary helps to short-cut this time and effort, while still benefiting from the learning.

Intermediaries

If an intermediary is involved, they typically require extensive teams to cover, depending on the role of the intermediary:

- decision-making about any questions or issues as they arise; ideally this should be conducted by one leader and a board or steering committee;
- business and technology experts, who can assess the feasibility and innovation of proposed projects, and support data holders in their selection and preparation of suitable data to share;
- communication between data holders, data users and the intermediary, and any other third parties such as contractors that might be involved. These should be able to understand and 'translate' between different terminologies common in either area;
- recruitment of data holders, data users, or both. This will require sufficient staff; existing personal networks in either ecosystem will make this significantly easier;
- drawing up contracts, negotiating terms, defining and explaining rules for compliance within the project, all of which requires legal experts. Their expertise may also be required for due diligence checks;
- developing and conducting due diligence checks, and keeping track of information and documentation from all parties, requiring sufficient administrative staff;
- sufficient staff to produce any other outputs of the project that need to be produced by the intermediary themselves.

A note on financing

So how might this all be financed? Questions to address are whether and how intermediaries fund data users for their projects, data holders pay or invest into data users, or data users pay for access to the data.

Decisions regarding this are affected by why the data is being shared. If it is for corporate social responsibility reasons, the cost may be appropriately financed from the sustainability budget, and asking users to pay for access may undercut the purpose of sharing.

Shared data can be directly monetised, either through a marketplace such as Dawex (see case study on page 6), or by selling licences directly. This obviously means that while revenue is created up front, there is little opportunity to set the direction of, or share in the upside of, innovation with the data.

Another option is to invest in the data sharing and then benefit from the results, through mechanisms such as new products and services that become available, or increased efficiency of internal processes.

Lastly, participation in a programme such as Data Pitch does not completely remove costs, but certainly removes some of the direct costs of the development of the legal agreement, finding solution providers or supporting their work.

A note on scale

Given all the resources involved in sharing data, the scale of the data sharing relationship bears some consideration. If data sharing is intended as a one-off engagement, the benefit may be limited and the resources hard to justify; if on the other hand, the goal is to engage in such relationships regularly, or continuously with multiple different parties for different purposes, then the learning and preparation may be a very worthwhile investment.

As an absolute minimum, to grant access to any closed data, the purpose of the use of the data needs to be validated. The preparation, checks and negotiation that are required to do this are the main reasons that such projects currently do not scale well. This could be changed by either reducing the details to check, which might go along with increased risks, e.g. of data protection breaches; or standardising parts of the process, such as the contracts, which may then not be able to capture the complexity of the individual data sharing relationships. However, there is substantial progress towards automated management that will allow scale, such as Cisco's policy broker within the Manchester IoT CityVerve data hub,[28] or work at the University of Southampton is focusing on algorithmic policies that incorporate rules for the data into the data itself, such as who can use it and what other data sets it can be combined with.

Thinking about achieving scale might mean thinking bigger: Having a bigger or further reaching challenge for data users to respond to, so that more data users can address different aspects of it; making more data available so that there is variety in opportunities to address the challenge; iterating these challenges, so that not only the framework for a relationship, but also the learning derived in the teams that are involved in the process can be reused. Data sharing is not hard in itself; it is hard because sharing data to generate value through artificial intelligence or machine learning is a new concept, experience is limited, and so setting up a data sharing relationship goes along with a tremendous amount of organisational learning. Once this learning has happened, applying it to more and different scenarios will be easier. Building up that organisational knowledge costs time and resources; reusing it can make data sharing relationships scalable, and increase the return of necessary investment.

Intermediaries can be instrumental in building this knowledge. They can also achieve what individual data holders and data users may find more challenging: they can scale, providing matchmaking services between a multitude of different data holders and data users. Along with the matchmaking, the training and support they may provide can be scaled, as can the due diligence checks they conduct, especially if sensitive data is to be shared.

While some of these intermediaries, like Data Pitch, are currently funded publicly, business models could and should be developed to offer 'Data sharing as a Service'. This is often used in the short term, for example via datathons or hackathons, or projects such as the Alan Turing Institute's Data Study Groups (see case study on page 10), but a longer term solution, or one that had a broader remit than open innovation is also possible.

Institutional oversight could be another useful tool to enable scale. Currently, the only regulatory authorities involved in data sharing in the EU are the national data protection authorities, and the European Data Protection Supervisor. While their work is necessary and very valuable, their remit may not be broad enough: They supervise, but do not actively regulate. An EU-level regulatory agency, similar to the European Banking Authority, could oversee the use of data and define standards, which might then be validated in some automated fashion.

Key resources:

Towards a European Data Sharing Space (Lopez de Vallejo et al., 2019): Position paper outlining opportunities and challenges for data sharing spaces, and recommendations for their implementation.

http://www.bdva.eu/node/1277

[28] BT, 2017

GLOSSARY

This toolkit is intended to be accessible to users from a wide range of backgrounds.
We have avoided technical language, however, the following glossary may be of use.

- **Access Controls:** Security measures applied by a data holder or provider to any data user with which it proposes to share its data. These include placing terms and conditions on the use or reuse of the data, or allowing the data user access to the data only under some specified data environment.
- **Anonymisation:** Techniques for lowering the risk of identification of data subjects from data, typically by removing or aggregating data that would (help) identify data subjects, combined with other measures, such as adding noise.
- **Confidential Data:** A term from common law, to refer to data or personal information which is shared in confidence with another party, such as a lawyer, accountant or doctor, in order to allow the second party to act in their client's interest. Such information should not be shared with any third party except with the clear consent of the client; this, if it happens, is called a breach of confidence. Confidentiality agreements are often implicit, when confidentiality is a reasonable expectation of the client.
- **Data Controller:** A legal term from the GDPR, to refer to a person, company, or other body that determines the purpose and means of personal data processing (this can be determined alone, or jointly with another person/company/body). The Data Controller is responsible for what happens with the data, and held accountable for any breaches of data protection.
- **Data Environment:** The context in which data is held. Data environments are characterised by agents with access to the data, other datasets with which the data may come into contact, governance arrangements for the data, and infrastructure used to store it. Typically, a dataset will be stored in a range of different data environments. Data sharing usually involves moving the shared data from one environment into another.
- **Data Owner or Provider:** An entity that owns a dataset; this could, for example, be a company with sales data, or a GP with a patient database. For data sharing to take place, a data owner or provider must facilitate access to the data for a data user. Note that if the data owner facilitates such access to personal data, then this will be regulated by GDPR; if it is not personal data, then it won't be.
- **Data Processing:** A legal term from the GDPR, to refer to any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction of data.
- **Data Processor:** A legal term from the GDPR, meaning a natural or legal person, public authority, agency or other body which processes personal data on behalf of the Data Controller. A Data Processor is not necessarily also a Data Controller: It could be a third party that a Data Controller delegates the processing to, such as an external analyst.
- **Data Protection Impact Assessment (DPIA):** A risk assessment for the use of data that is mandatory under GDPR for processing that is likely to result in a high risk to individuals. It should describe the processing; assess its necessity and proportionality; assess risks to individuals; and identify measures to mitigate those risks.
- **Data Sharing:** The sharing of data between entities, typically for a specific purpose. This can happen between companies, or departments within an organisation. The data owner or provider provides a data user with access to some of its data. If the data is personal data, then the sharing will be regulated by GDPR.
- **Data Subject:** A legal term from GDPR, referring to a living person that is or can be identified through data. The term does not extend to institutions, organisations, or deceased individuals.
- **Data User:** A person or entity that uses data for their own purposes, for example for business, academic work or in government. Data users may transform data, for example by cleaning it up, merging it with other datasets, or feeding it into other systems.
- **Functional Anonymisation:** A risk management approach to anonymisation that accepts that whether data is anonymous or not is a function of the relationship between those data and their environment, and not a property of the data itself. Hence functional anonymisation goes beyond manipulation of the data, and encompasses manipulation of the data environment.
- **GDPR:** The General Data Protection Regulation; an EU regulation that came into force in May 2018. GDPR provides new definitions of terms such as data processing or anonymisation, and defines different bases on which data processing is allowed. It goes further than previous legislation in protecting data subjects, and as an EU regulation, unifies data protection across the EU, and thereby allows the flow of data across the single market.
- **Metadata:** Data that describes the properties of data. Metadata can be attached to a dataset, and can therefore be used to understand whether that dataset is of interest to potential users, without giving them access to the data itself. Of particular importance is metadata describing the provenance of data.
- **Open Data:** Data that is freely available on the internet, without access controls.
- **Provenance:** Metadata that gives a record of the inputs, entities, systems, and processes that were involved in the creation of data, providing a record of its origins.
- **Pseudonymisation:** Techniques involving the substitution of identifiers that are easily attributed to individuals with, eg, an ID number that is stored separately. Re-identification of the data is possible by reference to the original key; without the key, the data can be treated as anonymised.
- **Synthetic Data:** Data that has been created algorithmically rather than generated by real-world events. It is generally used to explore datasets before sharing, as a stand-in for test datasets of production or operational data, to validate models, and to train machine learning models.

'Want to get to grips with your data? Prepare to share it'

ABOUT THE AUTHORS

Gefion Thuermer

Gefion is a Research Fellow in the Web and Internet Science group at the University of Southampton. She works on the Data Pitch project, where she is responsible for producing reports about the key lessons learned from the experience of data sharing. She gained her PhD in Web Science with a thesis about the effects of the introduction of online participation processes in the Green Party Germany.

Johanna Walker

Johanna is a Senior Research Assistant at the University of Southampton and works on the Horizon 2020 Data Pitch, Interreg 2Seas Smart City Innovation Framework Implementation and European Data Portal projects. In her role in Data Pitch she led the negotiation of data sharing agreements and managed the governance arrangements. Along with Elena Simperl she authored a report on alternative models for data trusts to support the AI industry for the Office of AI. Johanna holds an MBA with Distinction from London Business School. Her PhD thesis is on the intersection of innovation, open data and data sharing.

Elena Simperl

Elena is Professor of Computer Science at the University of Southampton. She is a Director of the Web Science Institute, Director of the Southampton Data Science Academy and a Turing University Lead. She has led a number of influential AI and data sharing projects, including Data Pitch, QROWD, Data Stories and Data Market Services.

ACKNOWLEDGEMENTS

We would like to thank the European Commission for their funding and support of Data Pitch (Grant ID #732506).

We are also grateful to all the data holders and users, consortium partners, advisory board members, reviewers, industry experts, mentors, and team members who engaged in the project and helped us to make it a success. Lastly, we are grateful to all the people who agreed to be interviewed to support the development of this toolkit.

REFERENCES

BT (2017): Connecting Manchester: How BT's Internet of Things solutions became central to the CityVerve smart city project.

https://www.iot.bt.com/assets/documents/bt-city-verve-smart-city-report.pdf

BT (2017): Connecting Manchester: How BT's Internet of Things solutions became central to the CityVerve smart city project.

https://www.iot.bt.com/assets/documents/bt-city-verve-smart-city-report.pdf

Carnelley, P, Schwenk, H, Cattaneo, G, Micheletti, G, and Osimo, D. (2016). Europe's Data Marketplaces - Current Status and Future Perspectives. IDC

http://datalandscape.eu/data-driven-stories/europe's-data-marketplaces-–-current-status-and-future-perspectives

Curry, E (2016): The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. In: Cavanillas J, Curry E, Wahlster W (eds) New Horizons for a Data-Driven Economy. Springer, Cham

https://link.springer.com/chapter/10.1007/978-3-319-21569-3_3

European Commission (2017): Guidelines on Data Protection Impact Assessment (DPIA)

http://ec.europa.eu/newsroom/document.cfm?doc_id=47711

European Commission (2018): Towards a common European data space

https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0232&from=EN

European Commission (2019a): Building a European data economy

https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

European Commission (2019b): Guidance on private sector data sharing.

https://ec.europa.eu/digital-single-market/en/guidance-private-sector-data-sharing

European Commission (2019c): Regulation on the free flow of non-personal data.

https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data

European Statistical System (2017): Position paper on access to privately held data which are of public interest

https://ec.europa.eu/eurostat/documents/7330775/8463599/ESS+Position+Paper+on+Access+to+privately+held+data+final+-+Nov+2017.pdf/6ef6398f-6580-4731-86ab-9d9d015d15ae

Grossman, R. (2016) How Data Commons Are Changing the Way We Share Research Data and Make Discoveries: The Open Commons Consortium Perspective. NSF Data Science Seminar July 6, 2016

https://www.nsf.gov/attachments/139105/public/grossman-data-commons-NSF-16-v5p.pdf

Hall, W & Pesenti, J (2017): Growing the artificial intelligence industry in the UK.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

Hardinges, J (2018): What is a data trust. Open Data Institute

https://theodi.org/article/what-is-a-data-trust/

Hardinges, J, Wells, P, Blandford, A, Tennison, J and Scott, A (2019): Data trusts: lessons from three pilots. Open Data Institute.

https://theodi.org/?post_type=article&p=7888

Higson, C and Waltho, D (2009): Valuing information as an asset. White paper.

http://faculty.london.edu/chigson/research/InformationAsset.pdf

Johns Hopkins University, Center for Government Excellence: Data Inventory Guide.

https://labs.centerforgov.org/data-governance/data-inventory/

Khatri, V and Brown, C V (2010): Designing data governance. Communications of the ACM 53.1 (2010): 148-152.

https://doi.org/10.1145/1629175.1629210

Lopez de Vallejo, I, Scerri, S, Tuikka, T (eds) (2019): Towards a European Data Sharing Space. Brussels. BDVA

http://www.bdva.eu/node/1277

Noveck, B (2016): Data Collaboratives: Sharing Public Data in Private Hands for Social Good, Forbes

https://www.forbes.com/sites/bethsimonenoveck/2015/09/24/private-data-sharing-for-public-good/#3845d28651cd

O'Hara, K (2019): Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship. Web Science Institute White Paper

https://cdn.southampton.ac.uk/assets/imported/transforms/content-block/UsefulDownloads_Download/0326D18DCC9E4BD08816BB5F994FCA76/White%20Papers%20No1.pdf

OECD (2017): Enhanced Access to Data: Reconciling Risks and Benefits of Data re-use.

https://www.oecd.org/internet/ieconomy/expert-workshop-enhanced-access-to-data-reconciling-risks-and-benefits-of-data-re-use.htm

Open Data Institute (n.a.): Data Sharing Spectrum.

https://theodi.org/about-the-odi/the-data-spectrum/

Perez, L (2018) Why businesses aren't sharing more data. Open Data Institute

https://theodi.org/article/why-businesses-arent-sharing-more-data/

Schomm, F, Stahl, F and Vossen, G (2013): Marketplaces for data: an initial survey. ACM SIGMOD Record 42.1: 15-26.

https://www.cs.unibo.it/~montesi/CBD/Articoli/MarketPlaceForData.pdf

Stalla-Bourdillon, S and Knight, A (2017): Anonymous data v. personal data—a false debate: an EU perspective on anonymisation, pseudonymisation and personal data. Wisconsin International Law Journal, 34 (2), 284-322.

https://eprints.soton.ac.uk/400388/

Stalla-Bourdillon, S and Knight, A (2017): Legal and Privacy Toolkit v1.0

https://datapitch.eu/privacy-toolkit-v1/

Stalla-Bourdillon, S and Carmichael, L (2018): Legal and Privacy Toolkit v2.0

https://datapitch.eu/privacy-toolkit-v2/

Stalla-Bourdillon, S, Thuermer, G, Walker, J and Carmichael, L (2019) Data protection by design: building the foundations of trustworthy data sharing. In, Proceedings of Data for Policy Conference, 2019.

http://dx.doi.org/10.5281/zenodo.3079895

Thereaux, O, O'Donnell, F, Duarte, S, Ghani, B, Keller, J R, Tennison, J and Wells, P (2109): Anonymisation and open data: An introduction to managing the risk of re-identification. Open Data Institute

https://theodi.org/article/anonymisation-and-synthetic-data-towards-trustworthy-data/

Verhulst, S and Sangokoya, D, (2015): Data Collaboratives: Exchanging Data to Improve People's Lives

https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a

Verhulst, S, Young, A, and Srinivasan, P (n.a.): An Introduction to Data Collaboratives. Creating Public Value by Exchanging Data.

https://datacollaboratives.org/static/files/data-collaboratives-intro.pdf

Walker, J, Simperl, E, and Carr, L (2019): A Framework for Data Sharing for Open Innovation. In, Proceedings of the 17th International Open and User Innovation Conference, 2019.

Webb, M (2018): Creating the Energy Data Commons.

https://lo3energy.com/creating-the-energy-data-commons/

datapitch.eu