OPEN FOR FEEDBACK

WORK IN PROGRESS

Assessing risk when sharing data: a guide

February 2022

Open Data Institute

Contents

About

This guide has been researched and produced by the Open Data Institute, and published in February 2022. Its lead author is Deborah Yates, with contributions from Lisa Allen, Ed Evans, Jeni Tennison, Diana Szasz and Rachel Wilson. To share feedback by email or would like to get in touch about this topic, contact info@theodi.org.

The guide was inspired by, and draws on content from, existing openly licensed works:

- Department for Environment, Food & Rural Affairs, 'Open Data Risk Assessment'
- Geospatial Commission, 'Data Sharing Assessment'
- CAB International, 'Sharing agricultural data: managing risk to minimise harmful impacts
- UN Global Pulse, 'Risks, Harms and Benefits Assessment Checklist'
- Energy data taskforce, 'Modernising Energy Data: Open Data Triage'
- Centre for Data Ethics and Innovation, 'Privacy Enhancing Technologies adoption guide'
- Information Commissioner's Office, 'AI and data protection risk toolkit'
- Roche, 'Managing risks with personal data'

To share feedback in the comments, highlight the relevant piece of text and click the 'Add a comment' icon on the right-hand side of the page.

How can it be improved? We welcome suggestions from the community in the comments.

Purpose of this guide

This guide aims to help organisations identify, assess and manage risks related to sharing data that they hold.

From optimising supply chains and supporting innovation, to addressing sector challenges and delivering public services, we have seen that sharing data can generate benefits for companies, the economy, society and the environment.[1,2]

Yet when considering sharing data – whether on a one-to-one basis, with a group, or more widely – a common concern for organisations is in providing assurance to senior leaders that sharing a particular set of data will not generate negative impacts on reputation; compromise legal compliance or negatively affect their place in the market; or cause harm to society, the economy or the environment.

Without this assurance, people and organisations can find themselves in a situation where a lack of processes to consider and manage legitimate concerns about data practises can result in a reluctance to share data with them. This can limit value creation in the form of innovation, product and service development for the organisation wanting to share data, as well as for potential reusers of that data.

This guide seeks to provide early steps – prior to seeking legal counsel (if that is required) – to consider real and perceived risks in sharing data and to identify suitable mitigating actions. We include typical risk categories, key questions to consider and suggestions on how to minimise harm.

There is opportunity to save time and money by using this guide to identify any potential data-sharing issues early on, or indeed to identify where data is suitable for wider sharing; and to provide the assurance needed to share that data with confidence and provide the opportunities to create value from it.

Equally, sharing the results of a 'data-sharing risk assessment' – as facilitated by this guide – and being explicit about why data is not suitable for sharing could help to create insight and improve our understanding of data ecosystems and the barriers to effective data reuse.

The guide is **not legal advice** and should not be considered as such. It is intended to prompt consideration of real and perceived risks when sharing data and to identify suitable mitigating actions, with a view to encouraging wider data sharing.

[1] Open Data Institute (2020), 'Seven reasons why business should be sharing data', https://theodi.org/article/seven-reasons-why-businesses-should-be-sharing-data/
[2] Open Data Institute (2018), 'Using open data for public services', https://theodi.org/article/using-open-data-for-public-services-report-2/

Minimising risk at the point of data creation

This guide focuses on checking data prior to sharing. At the Open Data Institute our mission is to create an open and trustworthy data ecosystem, where data flows easily and works for everyone. To achieve this we believe one of the things organisations can do is to build in considerations around data sharing at the beginning of a project, instead of at the end. This section provides some pointers on how to do this.

Building in a presumption that data will need to be shared more widely in future gives opportunity to consider risks upfront when creating data, to minimise the risks and to save time when sharing data with others in the future.

Actions you can take to minimise risk at the point of data creation include;

- **Data minimisation.** Reducing the amount of personal data collected is one way of reducing risk when you want to share data. Consider sharing upfront when collecting data and ask yourself if you really need to collect or retain personal details? If the details aren't vital to your purpose, don't collect or retain them. For example, you may only need the first part of a postcode to achieve the same outcome and therefore, decide to delete the second part as it is not relevant for your purpose(s). For personal data, many data protection regulations make it illegal to process (collect, use, store, share) personal data without a valid lawful basis. Minimising the personal and sensitive data you collect will simplify the data sharing process, and reduce legal obligations.
- **Use openly licensed data.** When acquiring data from third parties, selecting data that comes with an open licence – including permissions for onward sharing – will increase the chance that your derived data set can be shared more widely. This can help to avoid possible intellectual property (IP) issues later on.
- **Consider data ethics.** Consider how social or personal influences might impact data collection or use, and the consequential effects of this when that data is shared. Bias can be conscious or unconscious and can result in under-representation of specific communities, which could impact them by giving an unfair advantage to others, or unfairly restricting access (for example, exclusive arrangements). Tools such as the Open Data Institute's Data Ethics Canvas and Tech Transformed's Consequence Scanning can help consider the intended and unintended consequences of data collection or related technologies.
- **Create metadata to help understand suitability for sharing**. Metadata is the important information (key facts and figures) about a dataset, giving it context and helping make it discoverable by internal and external users. Often metadata is stored as an annotated list of datasets in a data inventory. The context it provides can help users understand why data has been collected, what it contains, how it is managed and the ways it is suitable, or not suitable (for example, internal only), to be made available for others to use.

- **Increase data skills and knowledge.** Providing training – in data collection and handling; data protection requirements; bias in data collection; and commercial considerations – can help to avoid later problems when you want to share data. This might form part of annual information security training.
- **Validate data inputs.** Sometimes people and organisations input data using different approaches or formats, making it difficult to identify which fields might contain sensitive data or general data quality issues. Where possible:
    - Replace free-text fields with more validated data input fields, such as drop-down lists, to ensure input validation of fields.
    - Encourage data inputters to use a template approach when completing free-text fields, to ensure consistency.

The next section focuses on identifying risks in data that has already been created or collected. It provides helpful advice about how to manage these risks and about how to create value from data through sharing as widely as possible.[3]

[3] Bennett Institute for Public Policy (2020), 'Valuing data', https://www.bennettinstitute.cam.ac.uk/research/research-projects/valuing-data/

Managing risks when sharing data

For organisations intending to share data, risk management is often part of the process, and concerns over potential and actual risks can delay internal approval for sharing data.

Risks typically fall into four categories: legal and regulatory, ethical, reputational and commercial. The risk categories we include in this guide are generic, and applicable to all types of data (that is, not just personal data), regardless of domain, sector or geography.

Risk categories:

- **Legal and regulatory**: Perceived or actual risks of breaching data protection law, intellectual property rights, other regulatory requirements or legal contracts
- **Ethical**: Perceived or actual risk of enabling unethical data collection or use, or of directly impacting people and communities
- **Reputational**: Perceived or actual risk of suffering reputational damage from sharing or using data that breaches trust, or that reveals limitations in processes or analyses
- **Commercial**: Perceived or actual risk of losing competitive advantage in the market

Each category has its own section in this guide where we have included key questions to consider and suggested options to mitigate the harm to individuals, society, the environment and the economy; and to help facilitate data sharing with confidence. The questions are numbered to aid progress tracking and discussion, but the questions do not need to be answered sequentially.

Table 1 includes a summary of the key risks, associated questions and relevant mitigating actions for ease of reference. More detail and explanations of each risk category and relevant questions are included in each section of this guide.

The actual risks associated with a particular set of data will influence how widely the data can be shared. At the Open Data Institute, we want a world where data works for everyone. To attain this, we want to ensure that data gets to those who need it, particularly to address some of the biggest challenges of our time, set out in the UN Sustainable Development Goals.

The data spectrum (Figure 1) can help consider how widely data can or should be shared.[4] As data moves along the spectrum – from closed, to shared, to open –

[4] Open Data Institute, 'The Data Spectrum', https://theodi.org/about-the-odi/the-data-spectrum/

an increasing number of people will have access to it. As this happens, the potential to create value and impact for people, society, the economy and the environment rises.

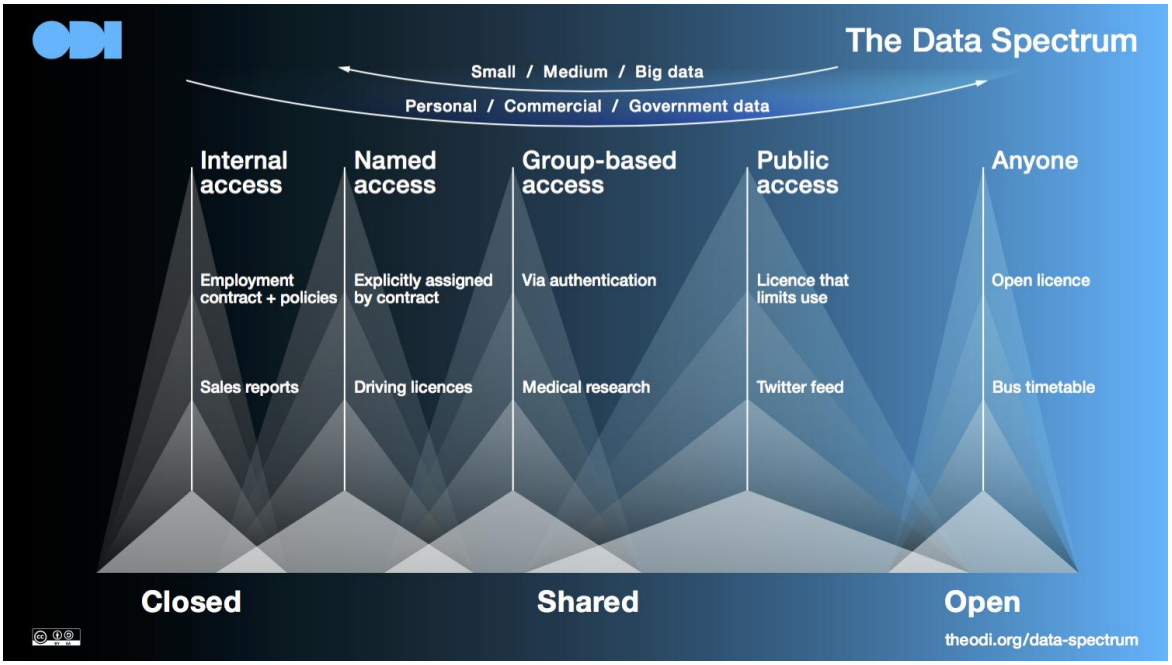

Figure 1: The Data Spectrum

Consider likelihood and severity of any impact

As with any risks, it is important to remember that while the impact resulting from a risk could be the same for different scenarios, for example, 'causing damage to the organisation's reputation', the likelihood and severity in each scenario could be significantly different, depending on the level of risk and the wider context.

Risk appetite is likely to vary widely depending on the organisation's context. For example, the level of comfort around sharing data with known quality issues might vary depending on the topic of the data, or who the data is being shared with.

To assess the real risk of harm, we recommend evaluating the likelihood of the risk occurring together with the severity of the impact. A common method of doing this is to complete a risk matrix, like in Figure 2, and assign a risk score.

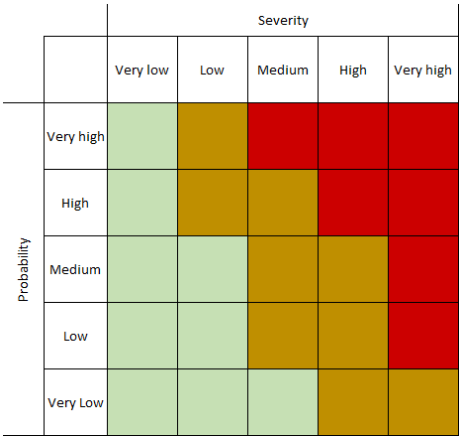

Figure 2: Risk matrix, David Vose (2018). Licensed for reuse: CC-BY-SA.

Regardless of the likelihood and severity, it may be possible to manage each risk using one, or a combination of actions. Table 1 illustrates this by indicating which mitigating actions are relevant for each risk area.

Table 1: Summary of risks & mitigation options

Risk category and specific risk	Questions	Mitigation options to minimise harm from sharing					
		Anonymisation	Synthetic data	Share under contract	Engage third-party data steward	Engage the community	Describe and document
Legal and regulatory Risks of breaching data protection law, and exposing personally identifying details	1. Does the data contain any personal data?	Y	Y	Y			
Legal and regulatory Risk of breaching legal contracts	2. Does the data contain third-party data?				Y		
	3. Do you have the legal permissions to share the data?				Y		
Legal and regulatory Risk of breaching other laws	4. Are there any other relevant legal considerations?	Y	Y	Y			
Ethical Risk of harm to society and the natural environment	5. Are there any relevant cultural considerations?	Y				Y	Y
	6. Is sharing the data likely to impact people or communities?	Y				Y	Y
	7. Will sharing the data impact the natural environment?	Y				Y	Y
Ethical Risk of harm to national security and staff	8. Does the data contain anything that could impact national security?	Y	Y	Y			
	9. Does the data contain anything that could impact the security of the organisation, or its staff?	Y	Y	Y			
Reputation	10. Will anyone be surprised by you holding, sharing or using this data?					Y	Y
	11. Is a data-quality caveat required?					Y	Y
	12. Are there any free-text or comment fields in the dataset?	Y					Y
Commercial	13. Does the data contain anything commercially sensitive?	Y	Y	Y			

Legal and regulatory risks

Legal and regulatory risks are the perceived or actual risks of breaching data protection law, intellectual property rights, other regulatory requirements or legal contracts when collecting, using or sharing data.

Within this category we have identified four key questions to consider.

1. Does the data contain any personal data?

Put simply, personal data can be defined as specific information about 'an identifiable person', such as name or location.

There are lots of different types of data about people, as outlined in figure 2.[5] Some types of personal data are more sensitive than others. Best-practice data-protection legislation defines sensitive personal information as 'special category' data and includes attributes such as race, ethnic origin, religious or philosophical beliefs, biometric data (where this is used for identification purposes) and health data.[6]

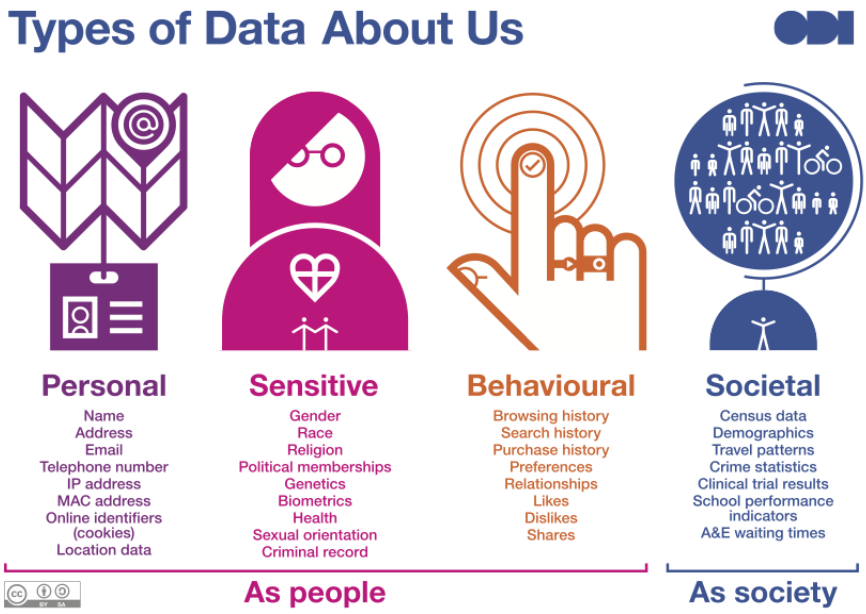

Figure 2: Types of data about people

Most countries will have different definitions and categories of personal data but generally speaking any data or information directly relating to an identifiable individual is personal. This includes images of a person, or group of people.

---

[5] Open Data Institute (2019), 'Types of Data about us', http://theodi.org/wp-content/uploads/2019/10/ODI-Types-of-Data-About-Us-graphic-20190916.png

[6] Information Commissioner's Office, 'Guide to the GDPR', https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/#scd1

Data protection regulations across the world are designed to minimise the risk of harmful impacts, while enabling personal data to be processed, that is, to be collected, accessed, used and shared. These regulations typically outline three key things:

1. The lawful basis for using and sharing personal data.
2. The rights of the data subject (the person the data is about).
3. Liabilities and penalties for breaching the regulations.

Examples of personal and sensitive personal data:

- Personal data: name, address, telephone number, IP address, location data, online identifiers (cookies)
- Special category (sensitive) personal data: age, gender, race, religion or belief, political affiliation, biometrics, disability, criminal record, health, sexual orientation, relationship status

If the data asset does contain personal data, there is a risk of identifying individuals and this could cause them harm and breach legal, regulatory or contractual obligations.

To help manage this, and share data as widely as possible, there are several common mitigation options available to minimise the risks of re-identification;

- **Anonymisation.** Anonymisation means processing data into a modified form that can be shared or published while significantly reducing the possibility of re-identifying individuals. Techniques include suppression of parts of the data, generalisation, randomisation and pseudonymisation.
- **Use synthetic data.** In some situations (for example research) it might be appropriate to share data that contains many of the statistical patterns of an original dataset, but eliminates identifying personal information. This is known as synthetic data, and it involves an automated process to make up (synthesise) data in a way that enables the same conclusions to be drawn from the data. This tutorial shows how to create a synthetic dataset.
- **Share the data under contract.** A contract with detailed, binding rules ensures all parties are clear on their legal obligations. Data sharing agreements can be useful when organisations of any kind are collecting, using or sharing data that is of a personal or sensitive nature.

2. Does the data contain third party data (data created by someone else)?

3. Do you have the legal permissions to share the data?

When an individual or an organisation puts intellectual effort into creating something, such as taking a photograph or collecting data, the law grants them specific rights of ownership over that work.

Different countries will have specific laws and definitions but generally speaking,

by default, the data creator holds exclusive rights to use the data, so that others must seek or be given the permission to use the data themselves.

Therefore it is important to review the terms under which you are using and sharing the data, to ensure you have the relevant permissions. These permissions are usually found in a licence accompanying the data, or in the contract (for example a data sharing agreement) setting out the terms under which data was provided.

The data you want to share might:

- be created and owned by you/ your organisation
- be completely licensed from someone else
- include an extract of content or data licensed from someone else
- be derived from the content or data licensed from someone else.

Examples of data typically sourced from third parties:

- Earth observation (for example from satellites), climate or weather, populations, political or administrative boundaries, transport networks or timetables, research, telecoms activity, social media behaviour.

If the data asset you want to share does contain third-party data or you are not sure if you have the legal permissions to share data, you can manage this risk by engaging the third-party steward.

- **Engaging the third party data steward.** Convening a conversation with the third-party data steward that provides the data can help to explore, understand and overcome any possible intellectual property (IP) issues and establish that onward use and sharing is permitted. If the data is not under an open licence and the steward places some restrictions on how the data can be used or shared, it may be possible to agree to share the data under a contract or licence that complies with the relevant permissions. Having these conversations can help to ensure legal risks around sharing data are minimised.

Examples of data licences and data sharing agreements:

- Creative Commons: Creative Commons include a range of licences with standard terms and different restrictions on use.
- Data-sharing agreements: These tend to be bespoke documents, outlining what data is being shared, for how long, and any restrictions on its use.
- Restrictive Licence Template: Developed as a part of the AusGOAL (Australian Government Open Access Licensing framework) for material that may contain personal or other confidential information. It may be used for other reasons, including material to be licensed under some form of limiting or restrictive condition (for example permission or ethics required,

a time limit on use, or contractual arrangements).

4. Are there any other legal or regulatory considerations relevant to this data?

There may be other legal or regulatory considerations from non-data-related legislation, or specific to your sector (for example the Equality Act 2010 and freedom of information requests) that will need consideration when sharing data.

This might include sector-specific legislation (for example financial institutions have particular duties and biometric data has particular limits). It might also include data licensing or intellectual property laws; or insights into data rights, for example, individual rights to data, rights for data creators, rights for governments and rights for citizens.

If you're not familiar with the requirements you may need to seek advice from specialists for your sector, area or on the type of data you want to share, to understand any obligations.

In addition, use of tools, like Consequence Scanning and the Data Ethics Canvas, can help to consider the intended and unintended consequences of data collection and surface wider legal, regulatory or ethical considerations. Consequence scanning provides an opportunity to reflect on risks that might not be immediately obvious.

Examples of other legal or regulatory considerations:

- Legal: Local laws on competition, intellectual property, digital economy, human rights, equalities act.
- Sector-specific legislation: Financial sector climate-related disclosures, oil and gas sector requirements around geophysical/seismic data, requests for environmental information.
- Policy: National data sharing and access policies or frameworks, requirements from international organisations that promote a specific type of data access, and sector or country codes of practice.

If there are other relevant legal or regulatory requirements you may still be able to share the data. To manage any risks, you could carry out the below processes:

- **Share the data under contract.** Data-sharing agreements can be useful when organisations of any kind are sharing data of a sensitive nature. A contract with detailed, binding rules ensures all parties are clear on their legal obligations.
- **Anonymise the data.** Anonymisation includes suppression of parts of the data, generalisation, randomisation and pseudonymisation. Redacting or

changing the data using these techniques could help to minimise risk from certain aspects of the data being shared.

- **Use synthetic data.** In some situations (for example research) it might be appropriate to share data that contains many of the statistical patterns of an original dataset. This is known as synthetic data, and it involves an automated process to make up (synthesise) data in a way that enables the same conclusions to be drawn from the data. This tutorial shows how to create a synthetic dataset.

Ethical risks

The increased use of data in recent decades prompts questions around issues of fairness, responsibility and accountability in relation to the use of data. It also triggers debate around whether existing legislation is fit to safeguard against harm to an individual's or group's privacy, welfare or safety, or to safeguard the environment.

Ethical risks are the perceived or actual risk of enabling unethical collection or uses of data, or directly impacting people, communities and the environment.

Increasingly, those collecting, sharing and working with data are exploring the ethical implications of their practices and, in some cases, being forced to confront those implications in the face of public criticism. Indeed, there is increasing pressure on organisations to report performance in relation to environmental sustainability and social responsibilities.[7]

Thinking about the ethical use of data is particularly relevant when insights drawn or decisions informed by data have the potential to directly or indirectly impact people and communities. When considering broader harmful impacts, think about the people the data is about, people impacted by its use, and the organisations using the data. For example, could use of this data result in decisions that discriminate against any groups or individuals, or impact their safety?

Bias can be conscious or unconscious and can result in under-representation of specific communities, which could impact them by giving an unfair advantage to others, or unfairly restricting access (for example, exclusive arrangements), therefore it is important to consider how data collection or use might be impacted by social or personal influences, and the potential consequential effects of this when that data is shared.

There is a huge body of research on the ethical risks of data use and sharing. Rather than repeat those here, the next three questions focus on data ethics, to ensure this nuanced area is considered so that data can be shared in ways that maximise positive impact, and minimise harm:

5. Are there any cultural considerations relevant to this data?

[7] European Banking Authority (2019), EBA ACTION PLAN ON SUSTAINABLE FINANCE'.

6. Is sharing the data likely to impact people or communities?
7. Will sharing the data impact the environment?

There are methodologies available that provide tangible and replicable ways to help answer these questions. Two examples are:

- Consequence scanning helps consider the intended and unintended consequences of data collection or related technologies. Consequence scanning provides an opportunity to reflect on risks that might not be immediately obvious, is part of responsible product development and can help provide structure to that thought process.
- The Data Ethics Canvas can help to identify and manage ethical considerations in projects involving data. It asks the user to consider 15 areas around data ethics – from bias in data sources to mitigating negative effects on people – to prompt critical thinking around how to collect, use and share data ethically. When sharing data it can be helpful to work through the canvas as a project team to promote understanding and debate around the foundation, intention and potential impact, as well as help identify the steps to ensure data is handled fairly.

Examples where sharing data may impact people, communities or the environment

- People and communities: Data flowing out of low- and middle-income countries (LMICs) to high-income countries via multinational corporations (also known as data colonialism, data nationalism, data repatriation), data about people in vulnerable circumstances, or dealing specifically with indigenous data governance and sovereignty
- Environment: Data that reveals the location of endangered species, data about natural resources.

If you think that sharing the data may directly impact – positively or negatively – on people, communities or the environment, then engagement and communication with those communities can really help to manage this. This could include the below actions:

- **Engaging the community.** Engaging those the data is about, or who might be impacted by it being shared, can help to validate or quash assumptions about the impacts, and identify actions (for example ensuring representation in the dataset or suppressing certain aspects) to reduce harm.
- **Properly describing the data.** Best practice data sharing includes publishing well-structured, high-quality documentation and metadata to accompany data releases. Documentation can help users to understand important context – such as gaps or biases in the data (for example. gender representation. Data documentation that is open and clear can help users to understand whether they might be able to use it. If you are sharing data, this guide on describing and documenting data well should help you to do this. There are also data-quality frameworks available (for example, UK government and ISO 19158) to help understand and

communicate quality of the data for certain purposes.

- **Anonymisation.** Suppressing certain aspects of the data, through anonymisation techniques could help to minimise risk from certain aspects of the data being shared, for example removing location fields to prevent endangered species from being tracked.

While we advocate for sharing data for the common good, we also want to ensure that data sharing doesn't happen when it would be harmful to do so. If, after exploring the mitigating options above, the risks of harm are too great then not sharing the data is always an option.

8. Does the data contain anything that, if made available, could impact national security?

National security, is broadly defined as the safety of a nation against threats such as terrorism, war, natural disaster, and could be put at risk through the release of data. This includes any data that could be used to cause actual harm, deprivation or fear of the same.

If the data asset includes details that you think may impact national security, you may want to **consider whether the data is already publicly available.** It may be that the elements of the data you are concerned about are already shared by the government, public or private sector organisation. For example, transport infrastructure is broadly available and used by many organisations for route finding. If this is the case, then sharing the same data within your dataset is unlikely to cause increased risk.

Examples of data that, if shared, could impact national security:

- Details of transport infrastructure, food production sites, nuclear sites, drinking water sources

9. Does the data contain anything that, if made available, could impact the security of the organisation, or it's staff?

As well as the physical security of an organisation (for example, protection from fire, flood, natural disasters, burglary), security includes protection against inadvertent loss of data through poor processes or system failures (often known as cyber security), revealing the location of sensitive physical assets or putting staff at risk.

Examples of data that could impact the security of an organisation or it's staff:

- Location of buildings or infrastructure where sensitive research is carried out or sensitive records are held, employee shift patterns, footfall, building layouts, details of security software.

To minimise risks to security of the organisation, or the nation, you may consider the following actions;

- **Anonymisation.** Anonymisation includes suppression of parts of the data, generalisation, randomisation and pseudonymisation – redacting or changing the data using these techniques could help to minimise risk from certain aspects of the data being shared.
- **Use synthetic data.** Synthetic data is an automated process to make up (synthesise) data that contains many of the statistical patterns of an original dataset. This should enable the same conclusions to be drawn from the data, but eliminate details that might impact national security. This tutorial shows how to create a synthetic dataset.
- **Share the data under contract.** A contract with detailed, binding rules ensures all parties are clear on their obligations. Contracts, such as data-sharing agreements, can be useful when organisations of any kind are sharing data that includes information of a sensitive nature.

Reputational risks

Reputational risks include the perceived or actual risk of suffering reputational damage from sharing or using data that breach others' trust, or in revealing limitations in processes or analyses.

The weight given to reputational risk when deciding whether to share data will depend on how important it is to your organisation to demonstrate trustworthiness with data and data practices (see the risk matrix in Figure 2). Reputational considerations might include the below:

- **Managing expectations around data use**. The importance of considering and informing people's expectations about data use was emphasised during the Covid-19 pandemic. The eighth Caldicott Principle reminds those using and sharing data of the idea of 'no surprises': the importance of considering and informing people's expectations to promote understanding and agreement about its uses.[8] By sharing data you are being more open about the kind of information your organisation accesses, uses and shares. However, if this is a surprise to people, this could affect your reputation, and the trust they have in your organisation, so this will need to be managed.
- **Data quality**. Quality of data can be a big concern for organisations, especially when it comes to sharing data. The level of quality required for each data set will vary depending on the purpose for which the data was collected, and will often consider several dimensions. For example, some decisions require up-to-date, complete and accurate data, whereas others are reliably informed by historic, aggregated data. Sharing data can help to improve its quality as people feedback on issues as they use it. Overall, being open and welcoming input and feedback is essential to help build a

[8] UK Govt, Health & Social Care (2020), 'No surprises', https://www.gov.uk/government/speeches/no-surprises

healthy, trusted ecosystem around the data, and can help to maintain reputation.

- **Free-text fields.** By definition free-text fields are not restricted in value, they are input fields that can contain long notes, so could easily contain information not fit for wider consumption (for example descriptions, notes of conversations, opinions, actions, feedback – which can be of a personal or sensitive nature). Free-text or comment fields can also make it difficult to aggregate data in a way that it can be reused, due to a lack of standardisation or validation.

With this in mind, we have identified three questions to consider to help manage reputation:

10. **Will anyone be surprised by you holding, sharing or using this data?**
11. **Is a data quality caveat required?**
12. **Are there any free-text or comment fields in the dataset?**

If you answered yes to any of the above, it doesn't mean you can't share the data for reputational reasons. You can build or maintain a trustworthy reputation through clear and open communications about data practices in general, and about the data your organisation collects, uses and shares.

You can build or maintain a trustworthy reputation by undertaking the processes as follows:

- **Engaging the community.** Clear and open communication can go a long way to help manage people's expectations. This might be about data practices – for example by publishing commitments, policies and approaches – being open about the types of data your organisation collects, uses and shares, and why (particularly for personal data) or direct engagement and events with key stakeholders.
- **Properly describing and documenting the data.** Best-practice data sharing includes publishing well-structured, high-quality documentation and metadata to accompany data releases. Documentation can help users to understand important context – such as quality, when the data was collected, update schedules and limitations in collection or accuracy – to guide their use of the data. Data documentation that is open and clear can help users to understand whether they might be able to use it, can help to manage concerns around mis-use of data and help manage expectations around the data that may lead to reputational concerns. If you are sharing data, this guide on describing and documenting data well should help you to do this. If you are sharing a computer model, this guide on documenting and sharing models should help. There are also data quality frameworks available (for example UK government and ISO 19158) to help understand and communicate quality of the data for certain purposes.
- **Anonymisation.** Suppressing or redacting certain aspects of the data – like free text fields – through anonymisation techniques could help to minimise risk that could come from these fields being shared.

Commercial risks

Commercial considerations are likely to be a driving factor for many organisations when it comes to sharing data. Considerations will include the perceived or actual risk of losing competitive advantage in the market, and/or the impact on commercial income.

However, sharing data can generate benefits for profitability, such as driving innovation, optimising supply chains, improving market reach, and generating new insights. Many businesses are embracing data sharing, and are seeing tangible benefits for their organisations and across their entire sectors.[9]

Here we have identified one key question that might influence the way you share data:

13. Does the data contain anything commercially sensitive?

Commercially sensitive information is any information that requires more careful handling to reduce harmful impacts; it may require restricted access. This could include considerations around commercial confidentiality, intellectual property and competitors gaining an advantage.

What is considered confidential is usually decided by the organisation who created or stewarded that information and will depend on the context.

Examples of commercially sensitive data:

- Related to competitive advantage: Unpublished copyrights, commercial records (for example customer lists, price lists, suppliers), research and development, unique processes (for example manufacturing techniques, mathematical formulae)
- About the organisation: Financial information, locations of properties, location of valuable assets, pollution level, wage gaps by protected characteristics, service complaints, lawsuits, affiliated organisations, employee records

If the data you want to share includes commercially sensitive information there are ways you can mitigate risks, and still share the data as widely as possible:

- **Anonymisation.** Anonymisation includes techniques such as suppression of parts of the data and generalisation to ensure the granularity of data is appropriate for the purpose, and to prevent sensitive details being revealed.
- **Use synthetic data.** An automated process to make up (synthesise) data that contains many of the statistical patterns of an original dataset. This should enable the same conclusions to be drawn from the data, but

[9] Open Data Institute (2020), 'Data sharing in the private sector', https://theodi.org/service/consultancy/business-data-sharing/

eliminate identifying commercially sensitive information. This tutorial shows how to create a synthetic dataset.

- **Share the data under contract.** A contract, such as a data-sharing agreement, can be useful when organisations, of any kind, are sharing data with embedded intellectual property rights, or commercially confidential data. A contract with detailed, binding rules ensures all parties are clear on their obligations. There are a wide range of approaches[10] for providing access to data with restricted permissions, including delegating data stewardship to a third party.

Next steps

We recognise that sharing data can be a nerve-racking thing. Now that you have worked through this guide we hope we have shown it is possible to manage real and perceived risks in ways that can still enable data to be shared as widely as possible and in ways that can help to realise very real benefits for people, organisations, the economy and the environment.

Now that you understand the risks with your particular dataset, you're good to go!

You could use your notes and answers from working through this guide to:

- support internal conversations and decisions about sharing the data
- obtain sign off for sharing from the relevant senior accountable person
- secure resources to put mitigations in place
- consider how you will be transparent about limitations of the data
- seek specialist legal advice, if there are still areas of concern
- (where you can see risks are low) publish data and contribute to the wider data ecosystem.

If you'd like further help or advice on this topic please do get in touch, contact info@theodi.org.

[10] (2019), Open Data Institute, 'Data Access Map', https://theodi.org/project/the-data-access-map/. (2020), Nesta, 'Data Sharing Toolkit', https://media.nesta.org.uk/documents/Data_sharing_toolkit.pdf.

Further resources

The list below brings together the resources referenced throughout this guide to help consider and manage risks with data sharing.

- The Data Spectrum: explaining the language of closed, shared and open data when considering how widely to share data.
- Personal data
    - Types of data about us
    - Openness principles for organisations handling personal data
    - The Information Commissioner's Office's personal data sharing code of practice
    - Introduction to anonymisation and the risks of re-identification
- Ethical considerations
    - Data Ethics Canvas**:** can help to identify and manage ethical considerations in projects involving data.
    - Consequence scanning helps consider the intended and unintended consequences of data collection or related technologies.
- Tutorial on how to create a synthetic dataset.
- Sharing data under contract
    - Publishers guide to data licensing
    - Guide to designing data sharing agreements
    - Data licences
        - Creative Commons standard licence templates
        - Restrictive Licence Template
- Documenting data
    - Describing and documenting data
    - UK government metadata best practice
    - Documenting and sharing models
- Data Quality
    - UK government data quality framework
    - ISO 19158: quality assurance of data supply