Data governance: from principles to practice

Civil society, volunteer data science skills, and open datasets

Workshop report
November 2020

THE ROYAL SOCIETY

Summary

There is a significant benefit to be gained from the better use of data, and civil society and volunteer groups can benefit greatly from the use of data that is open, accessible and meaningful. However, there are some important considerations relating to how civil society and volunteer groups gather the skills and infrastructures to make better use of data, and how they establish the systems to ensure that the collection and use of data is governed appropriately and collectively. Exploring the issues in this area puts into focus some of the main findings and recommendations from the Royal Society and British Academy report *Data Management and Use: Governance for the 21st Century*.

This report summarises the discussions at a workshop held in partnership with the Ada Lovelace Institute, the Alan Turing Institute, the British Academy, DataKind UK, the Leverhulme Centre for the Future of Intelligence and the Open Data Institute, on 12 March 2020. The workshop set out each organisation's different perspective on the opportunity for using data for the benefit of civil society, and the ways in which principles for the governance of data use can be put into practice in the voluntary sector.

Report structure

The principles set out in *Data Management and Use: Governance for the 21st Century* formed the structure of this workshop and therefore, this report, with each section exploring an aspect of their application (a detailed outline of the existing tensions and disconnects in data management and use, and the principles for data governance, are provided in Annex B).

Context

Learning from citizen science and environmental data: an opening reflection drawing on experience at the interface of academic research and citizen data science, to highlight some of the challenges in volunteer and civic society uses of data.

Principle 1

Transparent, inclusive and democratic decision-making about trade-offs: exploring the concept and practice of collaborative data maintenance – the process and data infrastructure by which organisations and communities share the responsibility and work to collect, maintain, govern and use data.

Principle 2

Individual and collective rights and interests: exploring data practices and social value, considering how data governance can protect both individual rights, goods and benefits, and collective rights, goods and benefits.

Principle 3

Seek out good practices and learn from success and failure: effective data governance should display a commitment to promoting good practice and embedding continuous learning as a way of improving practices and standards.

Principle 4

Enhance existing democratic governance: data management and use should support democratic processes, help enact democratic decisions and be subject to democratic oversight.

Conclusions and actions

This note concludes with some actions to promote data use by civil society, including supporting collaboration in the civil society community; developing guidance and case studies for organisations to learn from; providing support for technical literacy and in navigating the data space; and promoting inclusive dialogue. This note does not necessarily represent the views or positions of organisations or individuals who took part. A number of the sources referred to in the note can be found in Annex A: resources guide.

Full details of sources and locations for these resources are given in Annex A: resource guide and further reading on page 20.

Context

Open and shared datasets, pro bono data science skills, and civil society

This opening reflection draws on experience at the interface of academic research and citizen data science to highlight some of the challenges in volunteer and civic society uses of data.

Muki Haklay: learning from citizen science and environmental data

Four snapshots from the past 25 years, from the area of environmental information and the civic-societal use of it, can show multiple, persistent digital divides, and what we can learn from them at an organisation, community and individual level. This draws on a background of 30 years of looking at the creation of, public access to, and use of environmental information.

Access to Environmental Information

Nearly 25 years ago, in 1997, Friends of the Earth found themselves in a situation where, in the US, the Toxic Release inventory (TRI) – a database of what factories are releasing into the environment – had been open and available to the public since 1986. In 1997 a website (Scorecard) was created that shared this information openly on the web, but such information was not accessible in the UK. Susan Pipes and Lesley St James, two technologists from Friends of the Earth, received a donation of a SUN workstation, plus an Oracle Database, plus ArcInfo, geographic information system (GIS) software. With this free access to about £30,000 of resources, they set up a server providing information about the UK chemical release inventory, using a dataset that was passed to them from the Environment Agency. The FoE system allowed a user to put in a postcode to see what was going in a particular area via a website called Factory Watch. That changed people's ability to access information and, a year later, the Environment Agency released a website called What's In Your Backyard? enabling access to this information. This snapshot shows the unique ability and innovation of a civic-society organisation, where a web-mapping server – the like of which did not previously exist – was created by the organisation from scratch.

Who could use this information? In 1997 the size of the population that had access to the internet at a speed that allowed proper browsing of this website was just a few people in universities, potentially some journalists, and a few other people (in total, less than 9% of the population). It was about raising awareness, but there was still some distance to go in terms of understanding the information, in terms of the toxicity of the chemicals in the inventory.

Making data meaningful to civic society

In a second snapshot, going back to 1999, the national resource on air quality provided information enabling viewing of an automatic monitoring station. It was possible to click on each one of them and on each of the pollutants. However, the output was inherently a CSV (comma separated variables) file of values, which had to be made sense of. This continued to be updated, and by around 2008, it was possible to ask for a specific location and get information. But what does a "level of benzene" in the area mean to a member of the public? Again, what can society do with these numbers?

In the area of environmental information, data has been accessible and open for a very long time, but how has it led to actionable knowledge? Between 2000 and 2010, air quality was not a priority issue for civic society. By 2010 it became more prominent, but the data was not enough to be considered actionable information – information that can be used as a basis for action. When information such as air quality data was used in community settings, the feedback was that 'this is not community information in community language that we can understand.' The Aarhus Convention was introduced with an assumption that the issue critical to participation in environmental decision-making, is access to information, but my argument is that while air quality information has been open for 25 years, we have found that communities find it difficult to make sense of it.

A third example is to note the significance of the form in which data is available. Since the International Geophysical Year in 1957, earth scientists and others have been used to using digital data – the use of computers has been on the increase since then and it is rarely considered now, even in field settings. In terms of communities however, we can see that in 1986 the Breeding Bird Survey was still relying on paper forms, and there is a project that is currently being run by Open Air Laboratories which is also in paper form. So, in community settings, data is not necessarily digital. It is important therefore to be aware that not all data is born digital, and aware of the technical abilities of groups.

Civic society is made up not only of charities that are big and technically capable, but often of volunteers that are more interested in the issue than in the form of the data.

The final example is from my current European Research Council (ERC) project, Extreme Citizen Science, Analysis and Visualistion (ECSAnVis) which is about creating data collection and allowing any community, regardless of literacy, to carry out citizen science activities. In an example from the Masai Mara, the community of warriors and the tribes there are concerned about the impact of climate change. With the support of Professor Jacqueline McGlade, who was the Chief Scientist of the UN Environmental Programme, they are now collecting data about tree health. They created an icon-based app whereby they can collect information about 170 types of trees and record the situation in each one of them. They recorded 7,000 data points which are then used with AI to analyse them with remotely sensed data.

The issue of literacy is also technological literacy. We are working with people who have not used digital technology before, and surely have not used mapping information before. But we have discovered that aerial imagery is accessible to anyone, anywhere, in the sense that it can be understood by them. If you have high-resolution and detailed aerial information, we have done enough experiments and studies to know that even people who have never seen it before in a digital form can understand it.

Understanding digital data and consent for its use

Still, there are issues relating to working with non-literate people in remote communities to explain to them the nature and mutability of digital data. There are also questions of how to deal with data that does not belong to individual, but belongs to the community. GDPR and data protection law is about the individual, and does not necessarily translate into data ownership ideas in different communities.

The ECSAnVis project is focusing on creating a visualisation tool to provide a way for people who have not seen digital technology and are not familiar with the transferring of information, to understand how data collected on a device can appear on a server somewhere else. This is done in a way that it is understandable and supports a meaningful conversation. We have also used practices such as capturing informed consent on video, rather than paper and signature, ensuring that there can be discussion of consent in contexts where paper forms do not have meaning. Across my work, it has always been that the data does not belong to me; at best, I am a custodian of the data. Every time I use it, I need to ask for the community for their consent.

Access to technology and resources

Another aspect is the need for dedicating resources for digital updates. As someone who has been running a social enterprise for 10 years which has been using digital tools, the burden of rewriting code from scratch every five years is very heavy, and at least can be supported through access to research funding. What is this like for charities, that do not have research funding?

PCs were difficult to access in the 1990s. Today, in order to access data science, charities need cloud servers – but what is the financial cost of that and what skills are needed to set them up? Using GIS is challenging – though made easier by recent apps, and open data can now be downloaded. But when working with marginalised communities, it is important to ensure that they don't have to pay for their own data access to use apps and data. And access to open data is not guaranteed – the GEOTHINK consortium in Canada demonstrated that some governments are closing down certain open services.

Demographics and data skills

When the issue is simply providing data, there is quite a big group of people who can access the data and use it if it is understandable to them. When the issue is use of a system, you require understanding about how to use the system, which, by definition, reduces the number of people doing that. When you get all the way up to creating new systems or setting up proper data-collection systems that will work on mobile devices – which requires specialist data science – the number of people with the necessary skills is really small.

There is also a persistent issue with the digital exclusion of some demographics, and a generational divide. There are people in their fifties and sixties, who are the community activists and would like to use computing and data, but may not have the skills to navigate existing tech. There is an untapped potential in the explosion of the number of people with qualifications from higher education, including those skilled in data science. But younger tech developers may not always know enough about user-centred design to contribute products that work for a diverse range of users.

Roles for intermediaries

Challenges in accessing data science skills, and in communities not knowing what to look for in data, can be addressed by using intermediaries. Mapping for Change is an intermediary set up to provide this ability to make sense of and access community information. Civic society is not expected to go to the Environment Agency website or to download the data directly, but they need to know who the intermediary is and how they find them.

But issues around making data meaningful to communities are ‘wicked’ and difficult to solve. We need to think about them as we go on in this area of citizen data science.

Discussion

The civil society sector, its challenges and opportunities

It is clear that there are many opportunities for the civil society sector to make use of data, but it also faces a wide range of issues. The table below summaries some opportunities and challenges. The rest of this workshop report sets out the ways that, by putting data governance principles into practice, they can be addressed.

ISSUE	OPPORTUNITIES	CHALLENGES
Data skills in civil society	There is a wealth of volunteers who are willing to support civil society	Civil society organisations need more skills in several areas, including technological skills, research method skills, understanding of ethics, and skills in community engagement and collaboration. The lack of skills and capacity within civil society could be addressed by involving more people, additional funding and more collaboration. It is important to ensure that willing volunteers are valued and cared for.
Access to data by civil society	Some data is open and available	There can be challenges to accessing data, even when data is available. Making data open alone is not enough, there needs to be further support to make it accessible, including support for developing skills. There are challenges relating to data quality and discoverability, with a lack of common standards and vocabularies relating to data. There are also questions around consent to use the data.
Defining data challenges	Enthusiasm, skills and influencing policy	Initiatives might garner enthusiasm and maybe even skills, but there is a challenge of problem definition that some organisations struggle to address. This is the first stage of outlining why the project is being carried out, and what purpose and outcome we are looking for, beyond the existence or availability of the data. This is about encouraging more of the strategic thinking that might be needed, ie the idea of strategy versus hacktivism. This can lead to opportunities to inform policy and to find and address gaps in services.
Securing long-term benefits	Efforts in different areas within civil society, of people working with data and building relationships around data can be really powerful, 360Giving and Open Data Manchester being an example	A challenge for civil society groups is ensuring that the open data is available, useful, and documented, allowing effective use and avoiding the risks of misinterpretation. What happens after a data-led civil society programme runs its course, is the data left to the side and never used again or incorporated into corporate systems? How can it be ensured that the organisations that are supposedly benefiting from these efforts do actually benefit? Key challenges to consider are: how can civil society enable people to empower themselves? How do civil society organisations ensure that their use of data really does empower poor and marginalised communities?

Transparent, inclusive and democratic decision-making about trade-offs

Open Data Institute: the example of collaborative data maintenance

Leigh Dodds, Rachel Wilson, Chris Thorpe and Julian Tait

The Royal Society and British Academy Data management and use report highlights that there are tensions to be navigated in the governance of data, for example between individual and collective rights, as discussed in the next section. It argues that if these trade-offs are to be navigated in a way that is transparent, then all of those affected should have real and effective opportunities to participate in making the choices. How does this apply in citizen data science?

The Open Data Institute (ODI) defines collaborative data maintenance as the process and data infrastructure by which organisations and communities share the responsibility and work to collect, maintain, govern and use data. Collaborative maintenance occurs in a wider context of "open culture", including open standards, open source code, and open data.

Open culture

Open standards are reusable agreements that can shape how we choose to collect, share and use data, available for anyone to access, use or share. They can be highly technical, for example focusing on file formats or data structures; or can be higher-level, such as standardised codes of practice or checklists. Open standards are most effective when all the organisations that might be impacted through adoption of a standard come together to help shape it. **Open source code** is when organisations or individuals work together to create reusable code and applications, resulting in mutual resources for helping deploy websites or data collection. **Open data** involves publishing data under an open licence, so that it can be accessed, used and shared by anyone for any purpose. Open data has historically been about increasing transparency and accountability, but recently the focus has shifted towards using open data to solve shared challenges and to address social, economic and environmental problems.

Collaborative maintenance

Collaborative data goes further than either open source code or open data, to emphasize collaboration across the data lifecycle of data collection, maintenance, governance and use. An example is OpenStreetMap and Humanitarian OpenStreetMap, a collaboratively produced map of the world developed by enthusiasts and large organisations such as Microsoft, Uber and Apple, as a shared collaborative system – with the community collectively involved in data collection, addressing data gaps, and maintaining data accuracy.

Research conducted by the ODI has explored the different ways in which communities can be involved in collaborative data maintenance. This ranges from deciding what data to collect to ensure relevance, fairness and equity; to sharing the maintenance and governance of the data in terms of data access, quality, and inclusive engagement; to working with open source code when developing the tools to support data collection, use and management. In so doing, collaborative data maintenance can engage communities and organisations in ways that address possible tensions around data governance.

To support the application of this framework in different scenarios, the ODI have produced a Collaborative Data Maintenance Guidebook (see resource guide). For example, if you are collecting data with the community, how do you manage quality when you might have people with very different skills and experience, and different data-collection devices contributing to that dataset? How might you be transparent about how data-quality is managed? And how might you build consensus across the community? To help navigate these issues, the language of patterns can be borrowed from architecture: each building is unique, but architects must manage similar tensions or challenges such as ensuring enough sunlight without causing excessive heat. A pattern catalogue is not overly prescriptive and instead allows people to quickly recognise that elements of a particular solution might work for them and the issue that they are working on, because of similarities in their context.

Leadership and data strategy

Data leadership is central to collaborative data maintenance. Within large organisations, there are people who will understand what the value of data is, but awareness may not necessarily exist at the top of those organisations. Within smaller organisations, there may be fear of uncovering processes that are not as robust as they might expect. To deter these issues, adopting a standard is key, and it should come from a leadership that is focused on making teams feel comfortable with the process of adopting more open practices. There is no expectation to be at a gold standard in the beginning of the implementation phase – it is more important to focus on the journey of an enhanced and comprehensive data strategy.

Promoting this culture depends on leadership across a number of stakeholders:

Leadership in government: The National Data Strategy should provide more training opportunities, as well as more opportunities for organisations to receive financial support to improve their data practices, through organisational change and access to training.

Leadership in big data companies: A lot of big data companies use their corporate social responsibility (CSR) funds to support exciting projects, however they could improve on how they channel and focus that funding and how they channel their in-house skills.

Leadership in charities: Trustees are key in charities, and digital trustee roles should be a standard asset and for it to be considered good practice to have a leader driving data strategy. It is also worth considering the feasibility of funding local Council for Voluntary Services (CVS) umbrella organisations to and support collaborative maintenance projects. Charities exist for their beneficiaries, and trustees need to be clear on the benefits offered by collaborative data maintenance. That includes the ability to show impact in relation to a charity's aims.

Managing organisational capacity and volunteers

Charities and voluntary organisations are often working at capacity with very limited time and resources. There is likely to be a need for a triaging of resources to ensure they are most appropriately spent, especially with limited volunteer resources. Strategic thinking is required to use technical volunteers and data resources collectively and in an impactful way.

Collaborative maintenance projects are often successful because they have made space to work with communities to develop solutions together. In doing so, the cost and burden is spread across organisations, sectors and networks. This type of multi-level engagement pays back dividends in the long-term.

There is also work to be done on how to make collaborative data management work for less exciting topics or projects. It is easy to get people involved in nature-conservation work for example, but there are other projects that people will not want to give up their time for, so finding a way to increase the appeal of these types of collaborative projects will be important to realise their potential impact.

Power dynamics

Adding to the time and funding constraints, the power imbalances relating to funding need more attention: who is funding a project and why? Collaborative maintenance may help address some of the tensions around co-production, equity and ownership. This requires deep and strategic thinking about the problems organisations are trying to solve with data collection, about issues around power dynamics in terms of who is guiding and leading these projects, and on the extent to which there is real opportunity for communities to get involved in shaping them. It is important to avoid a situation where those who are hosting data or providing the governance are also the ones with greater power. There is some ambition among those who are least advantaged in society to have a say or to be involved in the process of creating these datasets and this would start to even out the potential imbalances. It is important to involve such groups in governance, to avoid hierarchies and power asymmetries.

Individual and collective rights and interests

Data practices and social value

Another key principle for data governance from the report Data management and use was that data governance should offer meaningful and effective protection against both tangible and intangible harms, such as discriminatory treatment or exclusion from opportunities respectively. It should protect both individual rights, goods and benefits, such as health, and collective rights, goods and benefits, such as protection of the environment.

The Ada Lovelace Institute: The need to re-think data

Reema Patel, Jenny Brennan and Silvia Mollicchi

Rethinking Data is the Ada Lovelace Institute's largest-scale research and public engagement programme. It considers data systems, their complex and emergent properties, and the interaction between people and data systems. The challenge with data is its symbiotic relationship with people and society. In an emerging and complex system, it is essential to revisit the fundamental concepts that underpin how people think about data itself, and therefore it is necessary to rethink data through narratives, practices and regulation.

The opening position here is around the fact that data is never neutral; it is reflective of society. That means there is an interesting tension between data conceptualised as an objective description of reality in some way, and data as non-objective in the sense that it is created by people and often serves a certain purpose or delivers a certain outcome.

Rethinking data: key issues

There are issues that many have been articulated in different contexts around the **exploitation of data**. Data is often exploited through different ways, possibly through enclosure models, so it is gathered and then deployed and enclosed. Data enclosure may inhibit the ability to treat data as a public good, to achieve its social value, and even to fully understand how it can have social value.

The **rate of change in the emergent system** is another issue. There are political and administrative institutions, in the UK and beyond, who struggle to govern data in a holistic way and to acknowledge the central role it has in the modern world. The governments of France, Germany and the UK have all been working very rapidly to understand the emerging issues that are raised by the new use and governance of data.

There is a challenge around **agency** and over **how data is used**. If data is thought of as co-created by people, groups and society, it raises compelling questions about the nature of the relationship between the organisations that often use, deploy and apply data, and the individual to whom it relates. The nature of the conversations that we have about relationships between the NHS, patients going through NHS, and third party organisations such as DeepMind or Amazon are examples of the power asymmetries that may exist in the use of data.

Shaping a new future: key concepts

To shape future relationships between people, communities, data and AI, some key concepts need to be well understood. These include the following ideas and functions.

The case for the **social value** for data needs to be made and emphasised. It is essential to identify and recognise that **injustices** exist in terms of the governance and use of data. There are asymmetries of power, so how should they be tackled to have a more inclusive conversation about the benefits of data for everyone?

Data stewardship is a necessary and invaluable function in terms of the relationships between organisations holding data and the data subject, and in terms of the rights and responsibilities between them. What do those responsibilities look like and what does stewardship in the context of these rights and responsibilities look like?

Purpose-driven innovation is the idea whereby people talk about innovation as being responsible for generating outcomes that work for people in society. This is about ensuring the creation of the infrastructure for the effective use of data, and to enable innovation, is responsible, has legitimacy, is trustworthy and works for people and society.

Similarly, progress cannot be made unless the **regulatory frameworks** and the right kinds of **incentives** are in place in the system, and so it is about developing **data rights**. What is meant by 'data rights' is an interesting question, as a there is often focus on a very individualistic conception of data rights rather than a recognition of rights that may belong to groups of people.

Creating an inclusive language is important, along with understanding how the narratives around data impact on who can be involved in the discourse about data.

The notion of what good data practices look like across different contexts and spaces is worth considering. Looking at the EU's data strategy[1], it usefully distinguishes different 'data-sharing architectures', and the features that enable social value to be realised from data. Data trusts[2] are a good example, and there are tools such as Biobank, a model which that enables researchers to share biomedical data and also enables a wider group of people to access data for research purposes.

There is a lot to learn from the **different models of data access and data use** and different data infrastructures, which are the conditions in place that enable us to be able to maximise social value to be derived from data. This includes the legal frameworks, skills and tangible tools such as cloud and 5G, and considerations of who has the agency and control over that infrastructure.

Seek out good practices and learn from success and failure

The *Data management and use* report puts forward the principle that effective data governance should display a commitment to promoting good practice and embedding continuous learning as a way of improving practices and standards.

DataKind UK and the Ada Lovelace Institute:
Ways of working and creating social value – learning from examples
Reema Patel, Giselle Corey and Jenny Brennan

Sharing positive case studies

At DataKind UK, there are numerous case studies of data scientists and charities working together to create social value. One of DataKindUK's projects is a small foodbank in Huddersfield, which supports people with targeted intervention where necessary. The foodbank had an issue with thousands of people approaching them for food parcels, and only one support worker to provide advice, so they had to be extremely selective with time. Essentially, their question was: is there a way for individuals that come to the service to be directed to the support worker without waiting for ten visits, while not making the queue infinitely long? Data enabled prediction of the likelihood of an individual becoming dependent on the foodbank. Individuals with the most pressing queries were able to reach the front of the queue, and over time, they saw a decrease in the number of times someone returned – one of many indicators of success.

The Ada Lovelace Institute, Understanding Patient Data, the Office for Life Sciences and NHS England worked on a project highlighting the foundations of fairness for health-data partnerships through public engagement and deliberation, drawing upon citizen juries across England. Operational data in the NHS has value as it enables the prediction of pressures and challenges, for example, data can be used to predict how many people might access A&E at any given point in time. This example shows that individuals do not have to make major sacrifices, for example, on their privacy, for the NHS to benefit from being able to predict and prevent the capacity issues in the system. It also raises the possibility of anticipating when more resources are going to be needed, to protect capacity. Essentially, it is a great example of a data-driven intervention, that could potentially transform the way the NHS works.

The challenge of transparency and sharing lessons learnt

A challenge often raised is that it can be hard to always have perfect cases of best practice, and that hearing about cases where things go wrong is more common. Anxiety exists within non-profit communities about reporting, and creating examples, due to the risks in opening up practice to public scrutiny. An example of this from the private sector was Amazon identifying that its machine learning algorithm for assessing CVs generated bias. The response was immediately to question how such an algorithm could be created in the first place, as opposed to welcoming the transparency of being open about this problem and seeking an explanation of where the algorithm went wrong.

Sharing lessons learnt is hard to do in a culture and context where the media more readily report negative over positive case studies. Instead of being able to develop an acknowledgement of what went wrong and a viable plan to fix it, institutions focus on managing potential or actual backlash. Therefore, there is a lot of work to be done on creating safe spaces for learning about data use. The Financial Conduct Authority's regulatory sandboxes are a good example of recognising the need to balance creating a space for learning, while making sure that people are protected. There are lessons to be learnt from this as a sector, but also as a country.

It is valuable to reframe narratives about risks arising from the use of data, to the risks arising from not using data, in terms of opportunity-cost calculation. This means identifying the benefits that are not being secured for the communities by not taking the opportunities to use data. And while there are negative reasons for focusing on bad case studies, the research community or the civil society community often emphasises when things go wrong in an attempt to alert others. The best response is to learn from different perspectives, to share experiences to amplify good over bad practice. This requires peer networking and collaborations that work across disciplines, where people with stronger data-science and technical skills are paired with those who have specific questions or other skills that can be used to influence outcomes.

For a lot of systems, particularly algorithmic systems, and particularly in the public-sector, it is very hard for people to appeal decisions or even to understand the decision making process. This is problematic and currently there no body exists that places a requirement of transparency on such systems, or which implements checks and balances on decision-making models. It is therefore worth considering ways that people can safely highlight where issues have gone wrong, similarly to the aviation industry, where people are encouraged to report problems without the threat of being sanctioned. This approach would put learning and a system of transparency at the centre.

The importance of context

Case studies rarely exhibit perfection on all possible criteria. For example, some case studies demonstrate good practice in terms of the outcomes that they realise, others in terms of the level of engagement with stakeholders. Case studies might also shed light on different facets such as power dynamics, the legislative elements around compliance, and rights such as privacy, human rights and equitable distribution.

Context is also important to consider, as what works in one context may not necessarily work in another. For example, a local authority may reasonably see it fit to use predictive techniques to identify individuals who might need help with housing support, but it would be seen as highly problematic if they used the same techniques to predict which families are most likely to abuse children or put them at risk of neglect.

It is important to identify which aspects can be scaled or promoted as being repeatable elsewhere and which aspects require taking the context into account. The question of scaling up is difficult, and the way it can be done practically is to, essentially, scale a framework rather than an answer. This can be done using problem-centric framing, for example, moving away from what data exists and the benefits or drawbacks from that dataset, to focus on the problems or challenges an organisation has and how they can be overcome with data driven approaches. Tools like Doteveryone's 'consequence scanning' can be used to address and communicate about data ethics.

However, what is almost more important is not an off-the-shelf product to consider what constitutes good data sharing but, in fact, whether the people who are thinking about the data sharing, the data use and the management understand that there are issues that are often quite specific to the context in which they operate.

Examples of common data pitfalls and how to learn from them

- Using complex data collection, when simple in-person observation would be more helpful (eg for understanding who uses a public park).
- Datasets may be used for benign purposes but a malign actor can also use similar datasets.
- There are examples where projects wait to obtain certain data or consent, to later realise that that data is not needed.
- Data users may be misled into thinking that the data tells them everything when it does not – it is often necessary to consider other sources.
- There are circumstances in which people might think data exists, but it comes from flawed legacy systems or incompatible systems.
- Believing data is truly anonymised in circumstances when it is not. For example free-text data can still give away information about an individual even after anonymisation techniques have been applied.

BOX 1

Learning from examples of good data-sharing and data-access practices

There are a few specific examples of projects and environments that help demonstrate good practice:

1. The government portal on ethnicity (see Annexe A: Resources Guide, footnote 8), or TfL's API (see Annexe A: Resources Guide, footnote 19): examples that highlight a need for simplicity rather than creating mid-level access that neither meet the needs of people who want to do relatively simple qualitative analysis nor those of computer scientists who want to do a quantitative analysis.
2. iNaturalist (see Annexe A: Resources Guide, footnote 12): a collaborative approach for sharing data, so that there are some quality controls about how data moves from the community into more structured authoritative datasets.
3. World Bank development datasets (see Annexe A: Resources Guide, see footnote 20) and the International Aid Transparency Initiative (see Annexe A: Resources Guide, footnote 13): datasets or platforms that increase transparency and support democracy or research. The Natural Resource Governance Institute's contract-transparency database also increases the transparency with regards to resource-sector contracts, in a similar way Transparency International does for lobbying in the UK or EU context.
4. The UK Data Archive (see Annexe A: Resources Guide, footnote 38): an architecture and infrastructure to help support data-sharing to enable a choice of data sharing options for sensitive data.
5. Sense About Science and Understanding Patient-Data (see Annexe A: Resources Guide, footnote 41): initiatives that aim to improve communication with the general public and citizens about how science is done, how data might be being shared, and the positive reasons supporting acceptance of data being used.

Enhance existing democratic governance

The British Academy

The Royal Society and British Academy joint report Data management and use further argued that data management and use should support democratic processes, help enact democratic decisions and be subject to democratic oversight. Rights relating to 'ownership' need to be understood in the context of data use, based on a rigorous analysis of the concept of data ownership. Related to owning data is the act of sharing data, and in the context of the civil society domain, data trusts are an important mechanism for well-governed sharing of data.

Critical perspectives on data ownership

Hannah Knox

Ownership and metaphor

There are two issues that anthropologists tend to look at when considering the question of ownership: first, the link that is often made between producing or making something and owning it; and second, the observation that ownership often becomes contentious at moments of exchange. Current solutions to problems of data ownership might be inadequate solutions to an ill-defined problem and there is still more work to be done on working out what the issue is (see Annex A: Resources Guide, footnote 62).

When people use the metaphor of data as oil, they are evoking a particular frame: data is a commodity and something that can be owned, sold or traded. However, metaphors are continuously created for data, but they do not work perfectly. This is perhaps because data describes something else and is already an analogy. Perhaps many of the conversations around data are not even necessarily about data; they could be about control, power, or access to resources. This opens up the possibility that we might be talking past each other. Still, metaphors do have a role in terms of helping people understand abstract concepts. The interesting question is not so much what the right metaphor for data is but what these different metaphors are doing in terms of reflecting the nature of the socio-political debate about data.

Productive activity and ownership

The questions of data ownership seem to hinge on the question of who generated that data. People talk about 'my data', 'corporate data' or even 'public data', and these rest on the idea that the production process of making data gives the producer a claim to ownership. This relates to discussions around intellectual property, where legal mechanisms have been used to stabilise a link between the creation or production of knowledge or information and claims that then arise over the benefits derived from the use of that knowledge. Patents and intellectual property rights work to legally demarcate ownership in order to enable a type of financial or value recompense for the creator.

In approaching the question of ownership, from the anthropological perspective, the focus is less on trying to universalise or stabilise the relationship between productive activity and ownership, but instead trying to understand the processes by which that link is established. A lot of work that has been done on the case of cultural property rights can help reflect on questions of data ownership and balancing the rights and interests of different communities.

Concepts of cultural ownership

Many anthropologists have explored the question of what constitutes cultural property, and a lot of the discussion relates to questions about the repatriation of cultural objects or museum artefacts back to their places of origin. Where these demands for repatriation are concerned, indigenous communities invoke an idea of collective cultural ownership in order to establish property rights over these objects, but they can raise uncomfortable or untranslatable issues such as the relevance of cultural identity to claims to ownership. In contrast, in these processes of repatriation, museums have invoked other kinds of property relations in order to argue that they retain a claim over ownership of the objects. For example, they might claim that they made a payment to a community at the point of time of getting the object and, therefore, they have historically settled the relationship. Another claim is that the discovery of the object itself confers ownership, or that an object might have general scientific or social value for a broader community, which then conflicts with the question of the value that that object might have for a local community.

Similar to museum artefacts, new forms of data are also raising some perplexing questions about origins, the hidden conditions of their production, ambiguities over authorship and debates and discussions about relative responsibility. Discussions around cultural heritage in a project at UCL, that aimed at bringing data together, led to what is now known as the Museum of Data[3]. It is an ongoing project and has produced an interface in the early prototype stages. The format of the museum archive to create a database of a series of data objects has been used. The process of curation and collection to understand what a data object is, how it would be tagged, how it would be ordered and how it would be archived in a digital database was examined as a part of this. Entries to the catalogue included items such as a digital representation of the letter A, a smart meter, the first printout of the human genome held by the Wellcome Trust, and a snapshot of the original Google.com webpage shortly after its creation in 1998. However, a discussion-board posting by Satoshi Nakamoto outlining the principles of the Bitcoin cryptocurrency did not make it into the archive, due to questions about whether the object could archived. A key question was: who should the authorship of this email be attributed to? The message was written by Satoshi Nakamoto but the name is widely understood to be just a moniker for either an anonymous individual or a collective of individuals, not a person.

What these issues suggested is that what may be at stake in discussions about data ownership is the question of what different invocations of ownership mean in different settings and why they have power. The case of cultural property shows that the legal definition of property works with a very specific social and cultural understanding about individual authorship and about the tangibility and endurance of objects through time. However, when objects are not individually authored, when they are not located in a particular place, when they are intangible, when they do not endure through time, and when the agent that made it might be ambiguously an individual or collective, the question about ownership becomes intrinsically complicated and difficult to manage. Where open or public data is concerned, similar kinds of questions arise also.

Ownership and exchange

The issue is not to determine whether data can or should be owned but to resolve who can make a claim to data ownership, how such claims can be challenged, and what the implications might be in succeeding with a claim to ownership or not. To examine this further, the second part that anthropologists look at is around questions of ownership that are tackled through the practices of exchange.

Understanding what it means to own means understanding the conditions under what can or cannot be transferred to somebody else. To exchange is an act that requires what is being exchanged to become detached from the person or community who previously owned it, produced it or had it in order to pass it on to another person or body, who then becomes newly attached as the new owner. Anthropologists have categorised these two forms of exchange into a difference between the exchange of gifts and the exchange of commodities.

If ownership is a bit of a problematic term when it comes to data, thinking about what constitutes an appropriate exchange might be a more useful way of thinking about what is at stake. To assess this, we can compare with an example that is discussed by the anthropologist Marilyn Strathern, on the practice of organ donation.

Strathern pointed out that debates on organ donation hinge on whether the act of giving an organ should be an altruistic act or whether people should be monetarily compensated for what they give. On the one hand, some advocate a non-monetised form of organ donation and, in that context, the act of exchange is understood as giving a gift. However, there are also those who have received an organ and who sometimes say they want to repay the donor, but the donor then says that their consent is altruistic. This produces some tension. When there is a situation where somebody is giving away a part of their body as an altruistic gift, it often hides another feature of organ donation, which is termed as the idea that human donations enter the organ-procurement and distribution system altruistically but exit it commercially. This raises the question of whether people should be compensated for the giving of the organ in the first place.

A similar dynamic can be seen in discussions about data. The way in which data is thought about from an individualistic point of view often uses this language of gift exchange. People are asked to give consent for their data to be used, or to share their data with others. This invokes a language of reciprocity, for example, people may say that they recognise if they do not give their data, they will not have access to services like Google Maps, therefore articulating that there is a reasonable exchange at play. However, for those who see data exchange as problematic, it is the mismatch between this free gift of data on the one hand and the way that it is monetised on the other that is often at stake.

There are a few ideas that may be a step towards resolving some of these issues:

1. Shifting to a more distributed vision of ownership that may shift current practice.
2. Having opt-in options over opt-out when faced with accepting cookies can have an impact on whether people feel like they have ownership over the data, regardless of whether they do or not.
3. Having gift aid on data donations may help with data that people collect or process that is similar to restricted funding.

Critical perspectives on data trusts

Sylvie Delacroix

Data ownership and data rights

The idea of 'data ownership' has gained such currency in large part because of the common, intuitive association between 'ownership' and 'control'. People feel that they have lost control over something that is perceived to be increasingly important, and a common reflex is to therefore seek ownership. To understand why this intuitive association between ownership and control is particularly misleading when applied to data, it is helpful to compare data to a river.

If someone owns a piece of land, and that piece of land has a river that goes across it, owning a river does not mean that the landowner controls that river. There are many people both upstream and downstream who have non-exclusive rights; for instance, navigation rights, fishing rights, irrigation rights or milling rights. Owning a river, therefore, does not give you much control at all, but it does give you rights. When it comes to data, the river analogy helps make the point that what we really need to talk about are the rights that you have over your data, not so much the data itself.

Aside from being misleading (if one talks of ownership as a means to establish control), the concept of 'data ownership' is also a very poor answer to the problems at stake. Unlike in the river analogy, where people get to enjoy some rights over the river, most of the time data flows from people, which leaves them increasingly vulnerable. This raises a number of questions that are ill-suited to a framework that is defined in terms of ownership. A different type of framework is needed.

Bottom-Up Data Trusts

The framework we proposed, together with Neil Lawrence, is that of 'bottom-up data trusts', or 'personal data trusts'. There are three key aspects to this framework: (1) empowering groups to make choices about their data-reliant futures (2) addressing the vulnerabilities that stem from data sharing (3) putting in place a professional, intermediary layer between data subjects and data controllers.

1. Empowering groups:

At the end of the 19th century, in order to have the right to vote, you had to have a piece of land that you owned freehold. This led to the creation of land societies whereby hundreds of people would pool resources in order to buy one piece of land, shared amongst them and giving all of them the right to vote. Similarly, data trusts would now provide a way for people to gain collective agency and acquire a political and economic voice by pooling data rights.

This contrasts with the current situation, which allows decisions about our data-reliant futures to be made by a very small number of people. As data subjects, people live within a structure that could be compared to a feudal system, whereby there is access to infrastructure in return for the systematic exploitation of data by a few people.

2. Addressing vulnerabilities:

To address vulnerabilities associated with data exploitation, top-down regulation and further rights are needed, however it is a mistake to think the challenges raised by data can be tackled purely through top-down regulation. Most top-down regulation today is built around the notion of consent, which is widely acknowledged to be problematic and rarely meaningful when it comes to data transactions.

3. Professional, intermediary layer:

With data trusts, the proposal is to introduce a trusted intermediary layer between the data subjects and the data controllers, the data trustees. Data trustees have a fiduciary obligation of undivided loyalty, in order to effectively exercise data rights on the subjects' behalf. In the 21st century, just like we need doctors and lawyers, we also urgently need data trustees, because of the vulnerabilities at stake. The situational vulnerabilities that justify special obligations on the part of lawyers and doctors have a lot in common with those that stem from our leaking data on a daily basis.

Data need not be held centrally in the Trust: There are many ways of implementing data trusts. One approach could involve a centralised model. Another approach could be taken so that the data remains wherever it is, for example on the servers of Google or Facebook, and the job of the data trustee is to leverage data rights to, for instance, negotiate better terms and conditions or data-sharing agreements.

Challenging the one-size-fits all approach: Having different forms of data trusts would help overcome the currently rather one-size-fits-all approach to data governance. Concretely, in a matter of years, data trusts could provide a way of reversing the direction of consent. Instead of signing in, for instance, via Google, Facebook or Amazon sign-in when accessing a service, a data-trust portal would set the terms and conditions according to which your data can be collected and used. Repeatable terms and conditions for horizontal data-sharing represent the concept of data trusts that effectively promotes responsible data-sharing between companies by setting up horizontal data-sharing structures.

Choosing between different data-sharing structures: Different data-sharing structures are suited to different aims. When the suitability of a particular data-sharing framework is assessed, a value-based choice is being made. This value-based choice reflects how much weight the decision maker attributes to the need to protect people from certain vulnerabilities, or the need to promote societal goods, or the need to empower people to, for instance, monetise personal data or to monitor the quality of services.

Data trusts are not the answer to everything; they are one type of institution among others that are needed for effective data governance in the 21st century. What is key to enable both better data sharing and addressing vulnerabilities around data is to have bottom-up empowerment structures. Data trusts are only one form of such structure. Other data-sharing structures that do not rely on trust law include data co-ops, which could be a viable option to protect individual goods and, for example, enable individuals to monitor the quality of services. But data co-ops are less well-suited to taking into account the vulnerabilities at stake, as they rely on a contractual framework that does not come with the same kinds of safeguards as trust law does. In other words, different tools will provide more or less ability to address various priorities.

Introducing data trusts at the point of data collection

When policymakers must choose what governance structure will be best suited to their aims, they should first consider whether the data has already been collected. If it has not, that is where the proposed data trust structure is most helpful. Second, they should consider whether the data gives rise to rights. If there are no rights pertaining to the data, then data commons may be more sensible.

Ownership as a concept only makes sense and becomes salient if data is thought about merely as an economic-productivity tool, but data goes beyond that. There are many governance challenges associated with data management and use, and data-sharing structures such as data trusts constitute a set of helpful tools to approach these.

From principles to practice: actions for learning across civil society

Discussions at the workshop highlighted practical steps needed to enable civil society organisations to make the best use of data. The following is a summary of some of the actions needed.

Collaboration in the civil society community: Having a space for learning and sharing failures, a peer community for sharing what has been learned with other organisations and other colleagues.

Guidance and case studies: Clear guidance around licensing frameworks that will help people share data and the range of models from one-to-one sharing through collaborative approaches and data trusts; strong proofs of concepts of data trusts and distributed ownership in practice would be useful.

Direct support and technical literacy: There is a need for providing direct support for smaller organisations that are resource-constrained. There is an acceleration in data and digital capability, and for smaller organisations it is hard to keep up, so they need guidance and practical leadership.

Navigating the data space: support is needed to address problems with discoverability at the dataset level, in terms of finding data. There are also problems with discoverability at the organisational-networking level, where it is about getting people with complementary skillsets together. Good practice involves managing metadata and capturing the 'who, what, where and why' of data. The creators of any datasets should answer the following questions: 'Who collected this data, what is being collected, where, and why is it being collected?'. This will help other users understand what the initial research design was and what the purpose of it is.

Inclusive dialogue: The conversation needs to go beyond open data and include what data use and sharing means to the beneficiaries of civil society organisations, especially for people in situations of crisis. There is more to be done on accessibility and making sure that unheard voices are part of the conversation – not just between technologically literate people and the organisations using the data – but also about those who the data is about, service users and citizens. Bringing in voices who are affected or have been impacted by the use of data is critical. Engaging with communities is difficult and skills and training is needed to support this.

Annex A: resource guide and further reading

Complimenting data governance principle of sharing good practice, this resource guide comprises of useful resources, such as open datasets, publications, reports and useful guides with aim to share knowledge across the citizen data science community.

Case studies and examples – open/shared datasets

Awesome data (DataHub) is a collection on DataHub (publisher) presents collections of high quality datasets organized by topic. Topics include education, climate change and healthcare[4].

DataHub (Datopian). Datapian helps organizations of all sizes to develop solutions to manage their data and Datahub provides over thousands of datasets for free and includes a Premium Data Service for additional or customised data[5].

DFID spend publishing (open data IATI data format) (DFID) is a Development Tracker managed by the Department for International Development enabling the exploration of detailed information on international development projects funded by the UK Government[6].

Ethnicity facts and figures (UK government/Gov.uk) is a government data portal on the UK's different ethnic groups[7].

Get Information About Schools (Department of Education) is a register of schools and colleges in England where information on establishments, establishment groups (such as a local authority, trust or federation) or governors, can be searched and downloaded[8].

Government Outcomes Lab Projects Database (GoLab). GoLab sits in the Blatnavik School of Government (University of Oxford) and represents a ground-breaking example of research-to-practice innovation; this database holds their many international projects[9].

HOT (Humanitarian OpenStreetMap Team) is an international team dedicated to humanitarian action and community development through open mapping[10].

iNaturalist app (iNaturalist) is a mobile application that records encounters with other organisms and maintains life lists in the cloud. It also connects with experts who can identify the organisms that are observed and helps to build knowledge with other naturalists[11].

International Aid Transparency Initiative (IATI) is a global initiative to improve the transparency of development and humanitarian resources and their results to address poverty and crises. Their page includes both open data and support guides[12].

JRF data dashboard- useful poverty data (Joseph Rowntree Foundation) includes the latest UK poverty data and analysis[13].

Mozilla Common Voice (Mozilla) is a free database collecting a more diverse set of voices, to improve voice recognition. The project is supported by volunteers who record sample sentences with a microphone and review recordings of other users[14].

Open Contracting (Open Contracting Partnership) aims to make sure public contracts are open, fair and efficient and promotes the only international open standard on public contracts, the Open Contracting Data Standard, that is endorsed by the G20, G7 and major international organisations[15].

Open database of companies (opencorporates) is the largest open database of companies in the world and aims to make the world's company data open for all[16].

Open data for grantmaking (360giving) 360giving helps organisations to openly publish grants data and improve charitable giving, eg their COVID10: Grants tracker[17].

Transport for London Unified API (TfL) is the public TfL data for open data users to use in their own software and services[18].

World Bank Open Data (World Bank) provides free and open access to global development data[19].

Case studies and examples – volunteer data scientists

DataKindUK (DataKindUK) helps social change organisations use data science responsibly and is the most experienced provider of charity data science in the UK. They are supported by British organisations[20].

Examining the Black Box (Data Kind UK and the Ada Lovelace Institute) is a report that explores different approaches to algorithmic audits including bias audit and regulatory inspection. It is aimed at policymakers, and anyone who wants to understand better the options for better assessing the impact that algorithms are having[21].

Pro Bono OR case studies (OR Society). The Operational Research Society is the professional home of operational researchers and analysts in the UK and this page includes a tailored search by organisation type, problem type and technique, on various case studies and charities across sectors from health to crime, eg charities include Bloodwise and Crimestoppers[22].

The Humanitarian Data Exchange (OCHA) is an open platform for sharing data across crises and organisations and aims to make humanitarian data easy to find and use for analysis[23].

Practical toolkits

Collaborative Data (ODI) is a guidebook that helps people design and run projects that involve the collaborative maintenance of data[24].

Data and Public Services Toolkit (ODI) is a toolkit designed to help people designing and delivering public services[25].

Data collection tool for users with very little literacy (Sapelli) is an open-source project that facilitates data collection across language or literacy barriers through highly configurable icon-driven user interfaces[26].

Data management system for participatory mapping (GeoKey) provides server-side components to run participatory mapping projects[27].

DataPitch data sharing toolkit (DataPitch Innovation Programme) is a toolkit has been developed to help organisations that want to generate value by sharing data or facilitating data sharing and explains the concept, challenges, and processes to enable successful data sharing, alongside including resources and recommendations. It is an EU- funded open innovation programme bringing together corporate and public-sector organisations that have data with startups and SMEs that work with data[28].

Datapolis board game (ODI) is a board game about building services, websites, devices, apps, research, using closed and open data[29].

Data Safe Havens in the cloud (Turing Institute) aims to develop a policy and process framework for secure environments for productive data science research projects at a scale[30].

JOGLbeta Programs (challenges) (JOGL beta) provides toolkits on how to contribute to some data science initiatives and also on some COVID-19 related data challenges[31].

Open Standards for Data (ODI). This guidebook helps people and organisations create, develop and adopt open standards for data and it supports a variety of users, including policy leads, domain experts and technologists[32].

Policy design patterns that help you use data to create impact (ODI) aims to help government policy makers to see how data could be used to create impact[33].

QUiPP Synthetic data evaluation pipeline (Turing Institute). The Quantifying Utility and Preserving Privacy project aims to produce a framework to facilitate the creation of synthetic population data where the privacy of individuals is quantified[34].

Technical guidance resources (GoLab) includes guides on evaluating outcomes-based contracts, setting outcomes, awarding outcomes-based contracts, pricing outcomes, social impact bonds and on the Life Chances Fund projects[35].

Training and resources (Open Data Manchester) is the main repository for training and resources developed by Open Data Manchester and friends[36].

UK data archive (University of Essex) is home to the UK's largest collection of social, economic and population data for over 50 years, providing researchers with training, support and data access[37].

Citizen engagement and public engagement

African Open Science Platform: Strategy and Vision (CODATA, International Science Council) outlines the proposed governance, membership and management structure of the Platform, the approach to initial funding, immediate priorities and targets for 3 – 5 year horizons[38].

Colouring London (UCL) provides open statistical data about the city's buildings and the dynamic behaviour of the stock, they collate, collect, generate and verify over fifty types of data to visualise many of these datasets[39].

Data Science: A guide for society (Sense About Science) explains the language to help talk about data science and highlights the key questions to ask those people using data science as evidence in decision making[40].

ExCiteS group (UCL). Extreme Citizen Science (ExCiteS) is a situated, bottom-up practice that takes into account local needs, practices and culture and works with broad networks of people to design and build new devices and knowledge creation processes that can transform the world[41].

Discover Insight Group (Autistica) sits within the UK's autism research network that often promotes different studies to members and gather participants to take part in a range of research, the group gives feedback to researchers about how their research is put together, eg the use of language, data collection methods and accessibility[42].

H2020 QROWD (H2020) was a project that offered methods to perform cross-sectoral streaming Big Data integration including geographic, transport, meteorological, cross domain and news data, while capitalizing on human feedback channels[43].

Lectures without borders (sciednetwork) connects scientists with educational institutions and NGOs in various ways including online video lectures[44].

scistarter initiative (scistarter) helping collaboration between scientists and people who are curious or concerned and motivated to make a difference in fields including ecology and computer science. A citizen science project can involve one person or millions of people collaborating towards a common goal and public involvement is in data collection, analysis, or reporting[45].

Funding opportunities and other resources

Digital Fund (National Lottery Community Fund) is a fund that awarded £12.1m to 29 organisations and £500,000 for a support contract to work with them. They are currently doing discovery work to determine the needs of micro organisations and considering whether these are digital[46].

Healthier Lives Data Fund (Nesta) is a partnership between Nesta and the Scottish Government to invest in and support innovative digital technologies that make data available and useful to citizens to help them lead healthier, and more independent, lives. Grants are up to £30,000 over nine months and they are looking for bold people-facing projects that demonstrate innovative approaches to empowering Scottish citizens with data and information, and develop the potential of a new generation of data-driven digital technology[47].

Open Data in a Big Data World (International Science Council) is an accord proposing 12 principles to guide the practice and practitioners of open data, focused on the roles played by scientists, publishers, libraries and other stakeholders, and on technical requirements for open data. It also assesses the "boundaries of openness"[48].

Pro Bono OR (The Operational Research Society) is a scheme that connects volunteer analysts with good causes to donate their time and skills to help charities or other voluntary organisations facing difficult decisions or looking for improvement[49].

Statisticians for Society programme (Royal Statistical Society) provides information for third sector organisations looking for pro bono support, or statisticians/data scientists interested in volunteering[50].

Wellcome grant funding data (Wellcome Trust) is a grants browser for those with ideas in the areas Wellcome supports[51].

Analyses and horizon-scanning

Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance (Sylvie Delacroix and Neil Lawrence). This article proceeds from an analysis of the very particular type of vulnerability concomitant with our 'leaking' data on a daily basis, to show that data ownership is both unlikely and inadequate as an answer to the problems at stake. We also argue that the current construction of top-down regulatory constraints on contractual freedom is both necessary and insufficient. To address the particular type of vulnerability at stake, bottom-up empowerment structures are needed[52].

Building the tools for public services to secure better outcomes: Collaboration, Prevention, Innovation (GO Lab). The public debate around social impact bonds (SIBs) that touches on the extraordinary expectations and passionate opposition and which creates polarisation, can risk poor policy making. The GO Lab takes an agnostic stance and this report seeks to move beyond ideological debate to provide a constructive response based on robust academic evidence[53].

Citizen Centric Services for Smarter Cities (University of Southampton) argues we are seeing increasing recognition of data as the prime mover of smart cities, and the role of the citizen as both subject and object of data, has moved the focus back to this relationship. The framework presented in the publication for citizen-centred development of smart city services outlines a new, data-oriented co-production journey and offers suggestions for practical implementation through case studies, many of which the authors have been involved in[54].

Common Knowledge: Citizen-led data governance for better cities (Nesta). This report argues for radically new ideas about how the value of our personal information can be returned back to citizens that create that value in the first place, with a focus less on how money can be made from data, and more on how data can benefit society as a whole. The idea of 'data commons' – which offer us both a useful conceptual and a practical model for achieving better, more inclusive outcomes for data governance is explored[55].

Citizen Science (UCL) is a book that identifies and explains the role of citizen science within innovation in science and society, and as a vibrant and productive science-policy interface. The chapters consider the role of citizen science in the context of the wider agenda of open science and open innovation, and discuss progress towards responsible research and innovation, two of the most critical aspects of science today[56].

Counterfactual explanations without opening the black box: Automated decisions and the GDPR (Sandra Wachter, Brent Mittelstadt & Chris Russell). This paper presents the concept of unconditional counterfactual explanations as a novel type of explanation of automated decisions that overcomes many challenges associated with algorithmic interpretability and accountability. How this relates to GDPR and the 'black box' is explored and it concludes that unconditional counterfactual explanations can bridge the gap between the interests of data subjects and data controllers that otherwise acts as a barrier to a legally binding right to explanation[57].

Data management and use: governance in the 21st century (Royal Society and British Academy). In June 2017, the British Academy and the Royal Society published this major report on data governance in the UK. It argued that new uses of data create a series of pervasive tensions and so a one-size-fits-all approach to data governance across sectors, contexts, and categories of data would be inappropriate and ineffective. The report argued for an overarching principle that systems of data governance should promote human flourishing[58].

Data ownership, rights and controls (Royal Society, British Academy and techUK) is a note summarising the discussion and debate at the British Academy, Royal Society and techUK event on Data ownership, rights and controls: reaching a common understanding on 3 October 2018. Also included is a set of papers, which provide further explorations of data ownership, rights and controls[59].

Data protection by design: Building the foundations of trustworthy data sharing (University of Cambridge) is a journal article that explores data trust. requires further disambiguation from other facilitating structures such as data collaboratives. It argues that at a minimum, a process-based mechanism should have a trustworthiness-by-design approach at its core and that data protection by design should be a key component of such an approach[60].

Dynamics of data science skills (Royal Society) is a report that analyses the demand for professionals with highly specialist data expertise, which includes roles like data scientists, data engineers, statisticians, biostatisticians, economists and financial quantitative analysts[61].

***Efficient, but Effective?: Volunteer Engagement in Short-term Virtual Citizen Science Projects*.** This paper explores two short-term projects to understand how they influence participant engagement in the task and discussion elements of VCS. Descriptive statistics are calculated to characterise project participants and factors influencing volunteer task engagement and the effect this has on project outcomes are also looked at[62].

Protecting privacy in practice (Royal Society). This report provides a high-level overview of five current and promising privacy enhancing technologies (PETs) of a diverse nature, with their respective readiness levels and illustrative case studies from a range of sectors, with a view to inform in particular applied data science research and the digital strategies of government departments and businesses. This report also includes recommendations on how the UK could fully realise the potential of PETs and to allow their use on a greater scale[63].

Public views of machine learning (Ipsos MORI for the Royal Society). The Royal Society commissioned Ipsos MORI to carry out research into public knowledge of, and attitudes towards, machine learning, as part of a project aiming to increase awareness of this technology, demonstrate its potential, and highlight the opportunities and challenges it presents[64].

Rallying Together: Collaboration and public sector reform (GO Lab). This report by the GO Lab looks at how local authorities are joining forces with their local communities to overcome complex social problems. It explores 10 cases across the UK to understand why authorities are choosing to collaborate, as well as how they are doing so[65].

The UK data governance landscape explainer (Royal Society). This explainer provides an overview of many of the key UK organisations and structures currently responsible for data governance and an overview of UK organisations that advise on data governance. It focuses on cross-sector or cross-domain organisations and activities, as well as describing the roles of organisations that specialise in aspects of data governance[66].

Towards trusted data sharing: guidance and case studies (Royal Academy of Engineering). This report is aimed at organisations that have identified the opportunity to create value through sharing data and are considering setting up arrangements for data sharing. Through this project and other initiatives, the Academy aims to play a role in supporting better use of data in engineering sectors and in the broader economy[67].

Annex B: data governance tensions, disconnects and principles

In 2017, the Royal Society and British Academy's joint report Data management and use: governance in the 21st century argued that new uses of data create a series of pervasive tensions and disconnects that society will need to navigate if it is to use them to best effect. The report argued for an overarching principle that systems of data governance should promote human flourishing. Four high-level principles complement the need to promote human flourishing, as a framework for well-founded debate about the tensions inherent in data governance. These tensions, and the principles for navigating them, are set out below.

TABLE 1

Tensions and disconnects in data management and use.

TENSIONS	Using data to improve offerings...	...without limiting available information or choices.
	Promoting benefits fairly across society...	...while ensuring acceptable risk for individuals, communities and organisations.
	Promote innovation...	...while addressing societal needs and reflecting public interest.
DISCONNECT	Existing data governance concepts, such as 'privacy', 'ownership', and 'consent', are under strain.	Cause: the traditional data lifecycle (collection, processing, application) is no longer linear because of 'open networks of data' with interconnected and interdependent data lifecycles. Data collection and data use are harder to separate; non-sensitive data can hold sensitive insights; and data provenance can be unclear because of weak audit trails of meta-data and data trading and selling.

TABLE 2

Principles for data governance.

PRINCIPLE	ARGUMENT
Overarching principle: Promote human flourishing.	Human flourishing is multi-dimensional, dynamic and context-specific. It includes concepts such as wellbeing and the need for individuals and communities to prosper. At moments of contention, the principle should serve to reflect the fundamental tenet that society does not serve data but that data should be used to serve human communities.
Complementary principle: Protect individual and collective rights and interests.	Data governance should offer meaningful and effective protection against both tangible and intangible harms, such as discriminatory treatment or exclusion from opportunities respectively. It should protect both individual rights, goods and benefits, such as health, and collective rights, goods and benefits such as the environment.
Complementary principle: Trade-offs between data management and use are transparent, accountable, and inclusive.	Effective data governance must identify these competing considerations and balance them. To achieve this, decision-making must be multi-stakeholder and if necessary iterative. It should also be recognised that relevant expertise might come from both traditional and non-traditional perspectives, backgrounds and approaches. Finally, transparency alone is not sufficient: it must be accompanied by accountability.
Complementary principle: Enhance existing democratic governance.	Data governance and data use should support democratic processes, help enact democratic decisions and be subject to democratic oversight. There should be consistency and proportionality in governance frameworks and mechanisms, and appropriate balance between competing interests. Finally, enforcement powers and resources should be appropriate for achieving regulatory aims.
Complementary principle: Seek out good practice, and learn from success and failure.	Effective data governance should display a commitment to promoting good practice and embedding continuous learning as a way of improving practices and standards.

Annex C: workshop partners, Chairs, speakers and event team

This report is a summary of the workshop *AI and data governance from principles to practice: auto insurance* held at the Royal Society, 6 – 9 Carlton House Terrace, London, on 12 March 2020, in partnership with the Ada Lovelace Institute, the Alan Turing Institute, the British Academy, DataKind UK, the Leverhulme Centre for the Future of Intelligence and the Open Data Institute.

Workshop speakers

Co-Chairs

Professor Geoffrey Boulton FRS
Regius Professor of Geology Emeritus,
The University of Edinburgh

Dr Adrian Weller
Programme Director for Artificial Intelligence,
The Alan Turing Institute

Keynotes and panellists

Jenny Brennan
Researcher, Ada Lovelace Institute

Giselle Cory
Executive Director, DataKind UK

Professor Sylvie Delacroix
Professor of Law, University of Birmingham and Turing Fellow

Leigh Dodds
Director of Advisory, the Open Data Institute

Professor Muki Haklay
Professor of Geographic Information, UCL

Dr Hannah Knox
Director of the Centre for Digital Anthropology, UCL

Dr Natasha McCarthy
Head of Policy, The Royal Society

Reema Patel
Head of Public Engagement, Ada Lovelace Institute

Julian Tait
CEO, Open Data Manchester

Chris Thorpe
Head of Technology, CAST and The Catalyst

Rachel Wilson
Data Technologist, the Open Data Institute

Acknowledgements

DataKind UK

The Ada Lovelace Institute

The Alan Turing Institute

The British Academy

The Leverhulme Centre for the Future of Intelligence

The Open Data Institute

Event team

Anna Bradshaw
Senior Policy Adviser (Public Policy), The British Academy

Jenny Brennan
Researcher, Ada Lovelace Institute

Bronwen Butler
Policy Assistant, The British Academy

Dr Franck Fourniol
Senior Policy Advisor, The Royal Society

Zeireen Fuzurally
Project Co-ordinator, The Royal Society

Victoria Hallam
R&D Delivery Manager, The Open Data Institute

James Hannaford
Programme Co-ordinator, The Royal Society

Silvia Mollicchi
Researcher, Ada Lovelace Institute

Dr Mahlet Zimeta
Senior Policy Advisor, The Royal Society
(now Head of Policy, The Open Data Institute)

References

1. European Commission. (2020). *European Data Strategy*. [online] Available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en (accessed 13 October 2020).
2. Open Data Institute (ODI). 2018. *What is a data trust?* Available at: https://theodi.org/article/what-is-a-data-trust/ (accessed 21 October 2020)
3. University College London. *The Museum of Data*. Available here: https://museumofdata.org/ (accessed 19 October 2020)
4. Awesome Data. DataHub (see https://github.com/datasets/awesome-data, accessed 13 May 2020)
5. Datopian. DataHub (see https://datahub.io/, accessed 12 May 2020)
6. *What is the Development Tracker* (Department for International Development). Available here: https://devtracker.dfid.gov.uk/about/ (accessed 6 May 2020).
7. UK government. *Ethnicity facts and figures* (see https://www.ethnicity-facts-figures.service.gov.uk/, accessed 19 October 2020)
8. Department of Education. Get information about schools (see https://get-information-schools.service.gov.uk/, accessed 12 May 2020)
9. *Projects Database* (The Government Outcomes Lab). Available here: https://golab.bsg.ox.ac.uk/knowledge-bank/project-database/ (accessed 6 May 2020).
10. Humanitarian OpenStreetMap Team (see https://www.hotosm.org/, accessed 12 May 2020)
11. *iNaturalist* (2020). *iNaturalist App*. Available here: https://www.inaturalist.org/ (accessed 7 May 2020.
12. International Aid Transparency Initiative. Available here: https://iatistandard.org/en/ (accessed 19 October 2020)
13. *UK Poverty Statistics.* (Joseph Roundtree Foundation) [online] Available here: https://www.jrf.org.uk/data (accessed 11 May 2020).
14. *Common Voice (Mozilla)*. [online] Available here: https://voice.mozilla.org/en (accessed 11 May 2020)
15. *Open Contracting Partnership* (Open Contracting Partnership). [online] Available here https://www.open-contracting.org/ (accessed 11 May 2020)
16. *The Largest Open Database Of Companies In The World* (Opencorporates) [online] Available here: https://opencorporates.com/ (accessed 11 May 2020)
17. *360Giving* (360giving). Available here: https://www.threesixtygiving.org/ (accessed 6 May 2020).
18. Transport for London. Transport for London Unified API. Available here: https://api.tfl.gov.uk/ (accessed 19 October 2020)
19. World Bank. World Bank Open Data. Available here: https://data.worldbank.org/ (accessed 19 October 2020)
20. *Helping Charities With Data* (DataKind UK). Available here: https://datakind.org.uk/ (accessed 12 May 2020)
21. DataKind UK (2020) *Examining the Black Box*. Available here https://datakind.org.uk/examining-black-box/ (accessed 12 May 2020)
22. The OR Society. Case Studies. (see https://www.theorsociety.com/get-involved/pro-bono-or/case-studies/ accessed 12 May 2020)
23. OCHA. The Humanitarian Data Exchange. Available here: https://data.humdata.org/ (accessed 12 May 2020)
24. Open Data Institute. Collaborative Data Patterns. (see https://collaborative-data.theodi.org/ accessed 12 May 2020)
25. Open Data Institute. Data and Public Services Toolkit. (see https://theodi.org/service/tools-resources/data-and-public-services-toolkit, accessed 12 May 2020)
26. Sapelli (see http://www.sapelli.org/, accessed 12 May 2020)
27. GeoKey. Collect, share and discuss local knowledge (see https://geokey.org.uk/, accessed 12 May 2020)
28. Data Pitch (see https://datapitch.eu/datasharingtoolkit/, accessed 6 May 2020)
29. Open Data Institute. Datapolis board game (see https://theodi.org/service/tools-resources/datopolis, accessed 12 May 2020)
30. The Alan Turing Institute. Data safe havens in the cloud (see https://www.turing.ac.uk/research/research-projects/data-safe-havens-cloud, accessed 12 May 2020)
31. JOGLbeta (see https://app.jogl.io/search/challenges, accessed 12 May 2020)
32. Open Data Institute. Open Standards for Data (see https://standards.theodi.org/, accessed 12 May 2020)
33. Open Data Institute. Policy design patterns that help you use data to create impact (see https://theodi.org/article/policy-design-patterns-that-help-you-use-data-to-create-impact, accessed 12 May 2020)
34. The Alan Turing Institute. QUIPP-pipeline (see https://github.com/alan-turing-institute/QUIPP-pipeline, accessed 12 May 2020)
35. Government Outcomes LAB. Technical guidance (see https://golab.bsg.ox.ac.uk/toolkit/technical-guidance/, accessed 12 May 2020)
36. Open Data Manchester. training and resources (see https://github.com/OpenDataManchester/training-and-resources, accessed 12 May 2020)
37. University of Essex. UK data archive (see https://www.data-archive.ac.uk/, accessed 19 October 2020)
38. Committee on Data- International Science Council. 2018. African Open Science Platform: Strategy and Vision (see https://codata.org/initiatives/strategic-programme/african-open-science/, accessed 12 May 2020)
39. UCL. Colouring London (see https://colouringlondon.org/, accessed 12 May 2020)
40. Sense About Science. 2019. *Data Science: A Guide For Society*. Available at: https://senseaboutscience.org/wp-content/uploads/2019/06/SaS-DataScienceGuide-V8-SinglePages.pdf (accessed 14 October 2020).
41. UCL. Extreme Citizen Science (ExCiteS) (see https://www.geog.ucl.ac.uk/research/research-centres/excites, accessed 12 May 2020)
42. Austistica. Discover Insight Group (see https://www.autistica.org.uk/get-involved/discover-insight-group, accessed 12 May 2020)
43. CORDIS. H2020 QROWD – Because Big Data Integration is Humanly Possible (see https://cordis.europa.eu/project/id/732194, (accessed 12 May 2020)
44. Sciednetwork. 2018. Lectures Without Borders (see https://scied.network/2020/04/09/video-library/, accessed 12 May 2020)
45. Scistarter. Citizen Science (see https://scistarter.org/citizen-science, accessed 12 May 2020)
46. National Lottery Community Fund. Digital Fund (see https://www.tnlcommunityfund.org.uk/funding/programmes/digital-fund, (accessed 12 May 2020)
47. Nesta. Call for ideas- Healthier Lives Data Fund (see https://www.nesta.org.uk/project/healthier-lives-data-fund/call-for-ideas/, accessed 12 May 2020)
48. International Science Council (2017) Open Data in a Big Data World (see https://council.science/publications/open-data-in-a-big-data-world/, accessed 12 May 2020)
49. The Operational Research Society. Pro Bono OR (see https://www.theorsociety.com/get-involved/pro-bono-or/, accessed 12 May 2019)
50. Royal Statistical Society. Statisticians for Society (see https://rss.org.uk/membership/volunteering-and-promoting/statisticians-for-society-initiative/, accessed 12 May 2020)
51. Wellcome. Find a scheme (see https://wellcome.ac.uk/grant-funding/schemes, accessed 12 May 2020)
52. Delacroix, S. and Neil D Lawrence, N. Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance, *International Data Privacy Law*, Volume 9, Issue 4, November 2019, Pages 236–252, https://doi.org/10.1093/idpl/ipz014
53. Carter *et al*. (2018) Building the tools for public services to secure better outcomes: Collaboration, Prevention, Innovation. Government Outcomes Lab. Available here: https://golab.bsg.ox.ac.uk/knowledge-bank/resources/evidence-report/ (accessed 12 May 2020).
54. Walker *et al*. Citizen-Centre Services for Smarter Cities. Available here: https://www.southampton.ac.uk/wsi/news/2019/06/citizen-centric-paper.page (accessed 12 May 2020)

55. Bass, T. and Old, R., December 2019. Common Knowledge: Citizen-led data governance for better cities. European Commission (see https://media.nesta.org.uk/documents/DECODE_Common_Knowledge_Citizen_led_data_governance_for_better_cities_Jan_2020.pdf, accessed 13 May 2020)

56. Hecker *et al*. 2018. Citizen Science Innovation in Open Science, Society and Policy. UCL Press (see https://www.uclpress.co.uk/products/107613, accessed 13 May 2020)

57. Wachter, S., *et al*. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." Harvard Journal of Law and Technology, vol. 31, no. 2, Harvard Journal of Law and Technology, 2018, pp. 841–87.

58. *Data management and use: Governance in the 21st century* (The British Academy and the Royal Society, June 2017). Available here: https://royalsociety.org/topics-policy/projects/data-governance/ (accessed 12 May 2020)

59. *Data ownership, rights and controls: seminar report* (The British Academy, the Royal Society, and techUK, 3 October 2018). Available here: https://www.thebritishacademy.ac.uk/publications/data-ownership-rights-controls-seminar-report (accessed 14 April 2020)

60. Stalla-Bourdillon, S. *et al* (2020). Data protection by design: Building the foundations of trustworthy data sharing. *Data & Policy*, 2, E4. doi:10.1017/dap.2020.1

61. *Dynamics of Data science skills* (The Royal Society November 2019). Available here: https://royalsociety.org/topics-policy/projects/dynamics-of-data-science/ (accessed 12 May 2020)

62. Reeves, N.T., & Simperl, E (November 2019) *Efficient, but Effective?: Volunteer Engagement in Short-term Virtual Citizen Science Projects. Proceedings of the ACM on Human-Computer Interaction, article no.:177*.

63. *Protecting privacy in practice: the current use, developments and limits of Privacy Enhancing Technologies in data analysis* (The Royal Society, May 2019). Available here: https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/ (accessed 14 April 2020)

64. *Public views of machine learning* (Ipsos MORI for the Royal Society, April 2017). Available here: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf (accessed 14 April 2020)

65. Government Outcomes Lab (2019). Are we Rallying Together? Collaboration and public sector reform (see https://golab.bsg.ox.ac.uk/knowledge-bank/resources/are-we-rallying-together-collaboration-and-public-sector-reform/, accessed 12 May 2020)

66. *The UK Data Governance Landscape Explainer*. (The Royal Society, June 2020). Available here: https://royalsociety.org/-/media/policy/projects/data-governance/uk-data-governance-explainer.pdf [accessed 14 October 2020].

67. *Towards trusted data sharing: guidance and case studies* (Royal Academy of Engineering). Available here: http://www.raeng.org.uk/data-sharing (accessed 14 April 2020)

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- Promoting excellence in science
- Supporting international collaboration
- Demonstrating the importance of science to everyone

For further information
The Royal Society
6 – 9 Carlton House Terrace
London SW1Y 5AG

T +44 20 7451 2500
W royalsociety.org

Registered Charity No 207043
Issued: November 2020 DES7167