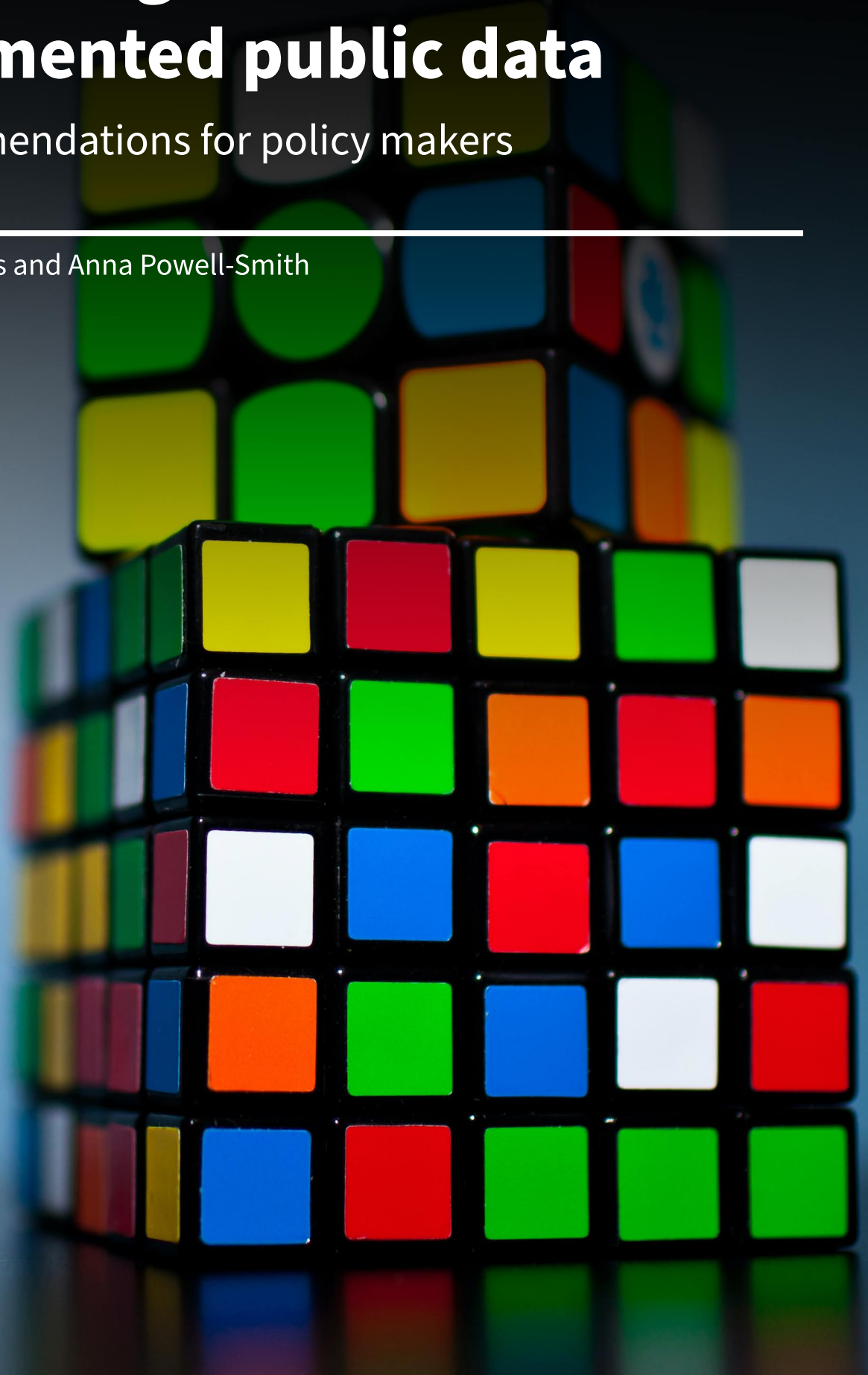


Unlocking the value of fragmented public data

Recommendations for policy makers

Alex Parsons and Anna Powell-Smith



Summary	3
About this project	4
About mySociety	4
About the Centre for Public Data	4
Copyright and licensing	4
Definition of terms	4
Why better data matters	5
The cost of fragmented public data	6
Our recommendations	7
Reflecting on previous waves of open data publishing	8
Enabling citizen (and civil society) scrutiny	8
Case study: Where the data isn't good enough to work with	9
Changing views of open data	9
How to create an effective publishing system	11
Create a collaborative (but compulsory) data standard	11
Meet data publishers where they are	12
Case study: 360Giving and spreadsheets	13
Work collaboratively	14
Make it a formal requirement to publish	14
Case study: Converging reporting definitions over time	15
Case study: Climate change duty reporting in Scotland and Isle of Man	15
If using a voluntary standard, be clear on the benefits to the data publisher	17
Case study: Food hygiene ratings	18
Create a central repository of the location of the published data and aggregate this data	19
Alternative (but not recommended) approaches	20
Discarded approach: Data publishers have a universal fixed URL system	20
Discarded approach: Data publishers have a dataset-specific fixed URL system	20
	1

Case study: Timely access to polling station data	21
Case study: Brownfield land registry	22
Provide support and assistance	23
Case study: Technical and non-technical barriers to local government publishing	24
Provide validation and publication tools	25
Case study: Investment in good tools means the data convener's job is easier in the long run	26
Conclusion	27
Acknowledgements	27

Summary

1. Fragmented public data is a problem that happens when many organisations are required to publish the same data, but not to a common standard or in a common location.
2. This is frustrating for everyone, reduces the economic value of data, and wastes taxpayers' money. Data users cannot easily use the data, policy makers do not see the impact they want, and publishers in public authorities are required to produce data without seeing clear results from their work.
3. This problem can be solved with increased coordination. The Government can provide this by requiring authorities to publish data to a common standard and in a common location, and providing support for a data convener. This would help businesses and other users unlock the full value of many public datasets that are currently underused, and any extra cost would be massively outweighed by the value created.
4. This report recommends three minimum features for a requirement to publish to be successful:
 - 4.1. A **collaborative (but compulsory) data standard** to agree the data and format that is expected.
 - 4.2. A **central repository** of the **location** of the published data, which is kept up to date with new releases of data.
 - 4.3. **Support from the data convener** to make publication simple - e.g. through validation and publication tools, coordinating returns, and technical support.
5. We recommend that:
 - 5.1. Whenever central government imposes duties on multiple public authorities to publish datasets in future, it should also provide the staff and budget to enable these features.
 - 5.2. The Central Data and Digital Office should publish official guidance covering the above.

About this project

About mySociety

mySociety is the charity behind UK civic services like TheyWorkForYou, WriteToThem, and WhatDoTheyKnow. We build open, digital solutions to help repower democracy, in the UK and around the world.

mySociety's climate programme is funded by Quadrature Climate Foundation and the National Lottery Community Fund.

About the Centre for Public Data

The Centre for Public Data is a new, non-partisan non-profit working for stronger public data. We work for stronger data collection and reporting across multiple policy areas, via advocacy, research, and original data analyses. We are non-partisan, and work with MPs and campaigners across the political spectrum to achieve change. Our funders for this work are the Mohn Westlake Foundation.

Copyright and licensing

This work is licensed under a Creative Commons Attribution 4.0 International Licence. (CC BY 4.0) It can be downloaded at <https://research.mysociety.org/publications/unlocking-fragmented-data>

Header image by [Olav Ahrens Røtne on Unsplash](#).

Definition of terms

Data publisher - organisation (or individual in an organisation) responsible for publishing a dataset.

Data convener - organisation responsible for ensuring the useful publication of a dataset, and potentially republishing it in more usable forms.

Data user - an individual or organisation other than the publisher, using the published data - this could be another public sector organisation, a company, a charity, private individuals, etc.

Why better data matters

Data that would help us make better decisions and build new businesses is too hard to access. This is a big problem, but one where a little work can unlock a lot of value. Publishing better public data would:

1. **Support policymaking:** Through local government, the NHS and the justice system, important information about how the country works is spread over hundreds of different public authorities. In the future, devolution and levelling-up plans involve moving new powers and responsibilities to different layers of government. Meeting the UK's net zero goals [requires important work and decisions at the local level](#), but crucially also means building new systems [of understanding what is happening](#). Improving how we do this is to the benefit of decision makers at all levels of government.
2. **Support businesses and startups:** Successful businesses can be built by analysing and adding value to public data - from procurement contracts to planning applications. These businesses create jobs and wealth, and make other UK businesses more efficient. Clean, available data reduces the barriers to entry for startups, and allows businesses to focus on adding value rather than basic data aggregation.
3. **Support public scrutiny of local and central government:** Greater transparency of public information is also a way of improving scrutiny of public authorities by citizens and civil society, ensuring that public spending is efficient and authorities are working as intended.

But publishing data on its own is not enough. Mandates that require no more than the publication of information can mean information is theoretically 'available' but in practice is stored across hundreds of PDFs, hidden on hundreds of different websites. This means that the work required to piece together the whole dataset from the outside is high, and as such most of the value of data is not realised.

How we handle publishing data is about the distribution of costs between publishers, intermediaries and users¹ - and where it is most effective to spend resources across the whole system. Our argument is that the value and impact of the data can be made significantly larger, if data publishers and conveners do only slightly more work upfront. Reducing the amount of duplicated work by intermediaries and users (or lowering the cost of making use of the data at all), makes it much easier for the data to actually be used, and have the impact and create the value that policymakers intend.

¹ Thanks to Connected by Data's Tim Davies for this formulation.

The cost of fragmented public data

In this report we are focused on the problem of what we call **fragmented public data**. Fragmented public data is when public authorities are each spending money to publish data independently, but their outputs are difficult to find and join together. This means that a lot of effort is going into creating data which cannot be used to its full potential.

The components of the fragmented public data problem are:

- Information is held collectively by many public authorities
- These authorities have a legal obligation to publish information
- There is no requirement to publish in a specific format
- There is no requirement to centrally link or deposit the data.

For example, every English local authority is required to publish all spending over £500 each month as part of the [Local Government Transparency Code](#). From [Adur](#) to [Wyre Forest](#), council officers are working hard to publish monthly spending.

In theory, this data should be being used by companies, researchers and journalists to provide insight into spending, spot fraud, and find opportunities to sell to councils.

But in practice, to use the data, you'll need to search all 333 English local authority websites each month, then import each spreadsheet into a central database – and this process will be very frustrating because the councils do not use a consistent format (either with each other, and sometimes within the same authority over time). As a result, not much has actually been done with all this data – and the grand promise that spending data would unleash an '[army of armchair auditors](#)' has largely failed to materialise.

This is a problem because most of the effort is already being spent on doing the job ineffectually. Councils do a lot of work to produce this data, and companies and analysts waste time fixing import scripts or crowdsourcing data, rather than creating new products or insights – and for many organisations, the skills and resources required to create national level datasets are beyond them.

This is not an isolated problem – there are many other examples of fragmented data. From [Assets of Community Value](#) to [election information](#) and [council land and property assets](#), data is often published in a way that is fragmented and hard to bring together. For many datasets, while individual disclosures are useful, the combined data is much more than the sum of its parts

because it allows real understanding of the picture across the whole country, and makes it easier to draw comparisons between different areas.

Across all these datasets the potential loss is huge – and just a bit of extra work could unlock huge amounts of the overall value of the data. We want to fix this for data that is already being published, and make sure that datasets in future are published in the best possible way.

Our recommendations

Whenever a government requires multiple organisations to publish the same data, it needs to set up some basic infrastructure. The minimum infrastructure is as follows:

1. A **collaborative (but required) data standard** to agree the data and format that is expected. This must be supported by:
 - A **collaborative relationship** with publishers to create a data standard that works for publishers and users, tools that are usable for publishers, and a process for listing and maintaining data locations.
 - This standard needs to be **required** as the form of data publication - not optional.
2. An online **central repository** of the location of the published data, so that data users can find it easily:
 - A list of **URL endpoints** for data is essential. This list should be maintained by the government (or other data convener).
 - This approach should assume URLs change for new versions of the dataset, and make it easy for data publishers to submit changes.
3. **Support from the data convener** to make publication simple and effective:
 - **Publication tools** to help organisations upload or publish their data in standard form. These can be minimal, like an Excel template or a simple web form; or more advanced like endpoints to upload JSON, depending on publishers' skills and resources.
 - **Data validation tools** to help publishers check that their data meets the standard. These don't have to be heavyweight: they can be built into an Excel template, or a web form; for more advanced publishers they could be command-line validators.
 - Enough **staff and money** for the data convener to agree the standard, work with publishers, maintain the tools and location registry, chase unsubmitted or invalid data, and provide a contact point for publishers and users.

Without these basics in place, decentralised data publishing initiatives will create fragmented data - costing taxpayers' money to publish, but rarely being used in practice.

Reflecting on previous waves of open data publishing

Some arguments for the benefit of publishing data will be familiar, and improving data publishing requires some engagement with why attempts to open up data in the 2010s have not worked as well as their advocates would have liked.

Enabling citizen (and civil society) scrutiny

Part of the explanation for this is the wider policy and political context of the period. Data publishing was seen as a potentially cheap way of improving accountability by giving ordinary citizens the information needed to hold authorities to account. In retrospect, greater support and coordination are required to properly make the data available in a form to enable outside scrutiny to have the largest impact.

The issue with the idea of ‘armchair auditors’ is not that they do not exist, but that they were thought about in the wrong way. [Research for Action](#)’s Audits of Local Government, and [Climate Emergency UK](#)’s volunteer-driven [climate scorecards](#) show that concerned citizens can be and are mobilised to use their time to support civic accountability. What these projects have in common is that they involve knowledge sharing and collaboration across the country, with volunteers themselves contributing a local or specialist focus.

Publication of accountability data will be most effective when it is in a form that groups like this can make use of, rather than seeing armchair auditors as isolated individuals. This kind of process could be made easier (and cheaper), if the government did the necessary work of collating the data, making it easier to discover, and enabling coordinating groups to create tools that can help volunteers cover more ground more rapidly.

Case study: Where the data isn't good enough to work with

The current data publishing landscape means it can be so hard to work with the data that useful projects just cannot happen.

Climate Emergency UK wanted to work out the average EPC rating of all council owned buildings as part of their work of compiling scorecards that marked [councils on their climate action](#).

In principle, all the data required to understand this is publicly available. All EPC certificates are available as open data: [councils are required to publish a register of assets](#). A joined up list of all assets could be simply cross-referenced with the EPC register for a sense of how different councils compare on energy efficiency. But the difficulty in obtaining the information from each council means it is prohibitively difficult to build a picture of energy efficiency of local authority buildings across the country.

One of the benefits of open data is that it can be joined in ways that were not originally anticipated. But where the data is fragmented, despite almost all of the work having happened, this benefit cannot be realised in practice.

Along similar lines, a journalist who works with local government data told us:

"I'm often in the business of trying to work out how many councils use certain companies (where this could suggest due diligence failures, or conflicts of interests). [...] By having all the material in slightly different formats, with different fields and using different services across hundreds of councils, and with no central data return, I'm often left having to search Google using advanced search to hope for hits by keyword, but some of the systems council use don't host the info on the surface internet. I tend to go to Spend Network who do a heroic attempt to collate some of this stuff, but that's only for some councils. And more transparency and cooperative councils in general have less to hide."

Bringing the data together is not just about creating one big dataset: making sure all authorities are publishing to the same standard would be fairer to authorities who are working above board, and makes it more difficult for information to be hidden through obscurity.

Changing views of open data

There is also room for reflection on how open data is, and has been, perceived. Some of the recommendations made by open data advocates in the 2010s made sense in terms of the context of the time (when there was a large culture clash between closed government data and open data

approaches), but today, systems need to build on past successes, and reflect on where previous tactics need to change.

The way that open data language has been ported from technical circles to government policy made open data seem more difficult than is helpful. The 2015 Local Government Transparency Code endorsed a '[five star](#)' system of data transparency:

Star rating	Description
One star	Available on the web (in any format), but with an open licence.
Two star	The same as for one star, plus available as machine-readable structured data
Three star	The same as two stars, but with a non-proprietary format (for instance CSV and XML, rather than Excel)
Four star	All of the above, plus use of open standards from the World Wide Web Consortium (such as RDF and SPARQL)
Five star	All of the above, plus the data is linked to other data to provide context.

The code recommended that data produced by local authorities should meet the three star format. This makes sense because the four and five star levels require significantly more technical skills, without providing any benefit to users (or fixing the fragmentation problem).

Our view is that this five star system is a bad way of understanding increasing technical difficulty and value to users in government data, which need not be related. We propose that a more relevant rating for fragmented data would look like this four level rating system:

Rating	Description
Bronze	Available on the web, in any format, with an open licence.
Silver	The same as for one star, plus available as machine-readable structured data.
Gold	The same as two stars, but available in an agreed format for the data (i.e. matching a schema, or using a standard template).
Diamond	All of the above, but the file, or a reference to the file, is also published in a central repository.

Here each stage enhances the benefit to users, and there is a clear justification why it is needed to solve a specific problem. While we believe that the most value is unlocked at the final stage, this rating can be used to grade existing datasets for how easy they are to create an aggregate dataset.

Our recommendations are in general more friendly to use of Excel and Excel file formats than other advice on open data. Older disputes about being locked into Excel's closed formats have in practice been mostly resolved by changes in default formats, but it remains common in advice and habit to

publish in CSV or ODS formats, despite the fact Excel's 'xlsx' format is an open standard (at least in the sense of not needing to use Office products to open) which has now been available for more than a decade. In practice, when given a choice, people publish and download .xlsx files, because Excel remains the main way people interact with spreadsheets. Meeting data publishers halfway is an important part of creating advice that can effectively improve the standard of published data.

How to create an effective publishing system

Through the rest of this report we explore these points in more detail, and through case studies illustrate current problems, and issues with alternative solutions that led us to our recommendations.

Our recommendations are primarily looking at the best way to publish a new dataset today, but have relevance on how existing fragmented datasets could be improved, or where there might be efficiencies in improving how organisations in general process their data.

A common thread among the people we talked to is that data publishers in public authorities have a variable amount of technical expertise, and generally many people who are publishing data may not view themselves as data professionals. The question is, does that mean lowering expectations about what can be achieved, or that improving skills (and possibly more defined data publishing job roles) is a vital part of better data publishing?

The answer in the long run is 'both', but focusing on better data publication sooner, we have concentrated on ways that technical processes need to adapt to work for the people who will be tasked with publishing it.

Create a collaborative (but compulsory) data standard

A data standard is an agreement about what data will be collected, and the form it will be collected in. We would put the emphasis on the "agreement" aspect of this rather than the technical form of the standard. The success of a data standard is ultimately about how people can both use it, and how the result meets needs.

In general, we recommend relatively lightweight standards, with support for publishers who lack high-level technical skills. Generally we envision that most data publishers will be interacting with Excel spreadsheets, simple online forms, or their own content management system (CMS).

Meet data publishers where they are

The technical form of a data standard can take the form of a defined 'schema'. This is a document which sets out how data is structured, which values are numbers, which values are words, validation rules to make sure the answer is valid, etc.

But while the tools used to manage data standards can become increasingly abstract and technical, this level of detail doesn't have to be what the data publisher sees. Once agreed, the data publisher's view of a schema may look like an Excel spreadsheet with columns that are decided in the standard. Validation rules can be expressed through a website that checks the data, or even directly in a web form the data publisher is filling out.

The design of data standards has to be sensitive to the realities of data publishing. While in some cases, data will be produced by a data team with good skills in data manipulation, in many cases the job of publishing a dataset will be given to the person managing the data as part of their existing job.

The level of technical expertise that should be assumed is a general working knowledge of Excel. This makes it important to design data standards that will work well for people who use Excel on a regular basis in their work. In practice, this means simplifying the data being collected so that it can fit in one table - even if it introduces redundancy. For instance, it might be better to include an organisation's name and address multiple times rather than including a separate table that maps an ID to organisation details. While the [360Giving standard](#) has some extra tables for additional information, all of the important information can be entered in one table.

Case study: 360Giving and spreadsheets

360Giving is a charity that supports organisations that make grants to publish in a common format. The aim is to make grant giving more transparent and make improvements across the sector.

To do this, they have to convince lots of different (governmental and non-governmental) organisations to publish in a common standard. Currently there are over 230 organisations using the format. The ways by which 360Giving supports and encourages publication, in the absence of requirements to publish in this format, are useful approaches for data release in general.

To get new organisations on board, 360Giving put effort into making publication as simple as possible — for example by letting people use familiar tools like Excel. While the data standard itself is a described JSON schema (a machine readable data format), few organisations actually publish as .json files. Instead, they publish their data in the form of spreadsheets (almost 85% as Excel .xlsx files, 3% as .ods, 3% as .csv and 3% as .json). The [Flatten Tool](#) can be used to convert structured json data into spreadsheets and back. While 360Giving's internal tooling expects .json files to run validation, externally, data publishers can work in spreadsheets.

The process is designed to work with data publishers of varying technical abilities. Data will often sit with a grant manager, who is not necessarily an especially technical person. The standard is constructed so that most information can be completed in one table (with a different grant in each row), but also can include additional tables for adding information to the dataset. As a way of helping data publishers translate from either existing spreadsheets, or outputs from other systems, 360Giving allow users to create [Conversion Tool spreadsheets](#). These use formulas to translate between the original output in Tab 1, merge with 360Giving specific identifiers in Tab 2, and create a 360Giving standard set of data in Tab 3, which can be copied out and published. As Katherine Duerden, 360Giving's Publishing and Support Manager put it, it's *"good to keep in mind that there's somebody whose job this is going to be - how do you keep them on board?"*

Work collaboratively

While it is important there is an agreed standard, this does not mean that standards need to be imposed from above. Instead, we recommend “enforced cooperation”, where data publishers and data users collectively have input and responsibility for forming how the process (that will eventually be required for all parties) will work.

Part of the reason for this is practical. People who work in public authorities are best placed to understand the data they hold, and how existing forms and processes may be adapted to publish it. While it is important that the end result is actually useful to users, publication processes will be more successful when they are clearly integrated into how authorities are already approaching work around a dataset or area of work.

There is a general philosophical point as well. Levelling up and local net zero approaches require both local flexibility and local decision making. Imposing data standards without significant input from local authorities is against the spirit of this. The goal of the central government should be to act as a supportive partner in the process of publishing data, ensuring it happens in a form that is useful to policy makers and the public, without making all decisions unilaterally.

Make it a formal requirement to publish

The other side of working collaboratively is that the agreed standard needs to be required and enforced. Requirements to publish information that are format-agnostic are not going to get the most value from the work data publishers are doing.

The benefits of a standard are often clear to those doing the data publishing. For instance, in [a consultation](#) about how Scottish public authorities should report on climate duties, public authorities were generally in favour of giving a specific form, with *“75% of all consultation respondents and 88% of organisation respondents [agreeing] that standardised reporting would improve the quality and consistency of climate change information reported by public sector major players”*. Consistency in process helps make clearer the case that the data is going to be used, and that it is a good use of data publishers' time to engage with the process. Even if publication is mandated, goodwill around engagement is useful.

In terms of the legal framework, we recommend a lightweight approach that defines the broad spirit and elements of data covered; a requirement to publish in a specific format, which defers to detail to a future document or code of practice to be published by a government department.

This is learning from the experience of publishing climate data in Scottish legislation. Writing exact data gathering requirements into legislation creates problems in adapting to feedback, and resolving ambiguities. It is better for legislation to create a reporting requirement, a sense of scope and reporting burden to be approved, but leave the fine detail outside of this process.

Case study: Converging reporting definitions over time

Part of the challenge of publishing data that is already held by public authorities is agreeing the definitions that sit behind which data is captured. Here both a collaborative and practical approach would be to accept broad definitions, and try to converge over time.

This was the approach of the Scottish Information Commissioner (OSIC). When trying to get authorities to report their FOI statistics into a central database, they found a lot of variation in how organisations were recording their Freedom of Information statistics. For instance, different organisations may regard different kinds of requests as falling under the Freedom of Information Act. Some smaller organisations recorded all requests as freedom of information; in other cases requests were only tracked when an exemption was being applied. Over time, OSIC has refined their guidance on what information should be held (e.g. they made an early clarification after the first collection for authorities to look at ‘requests being processed in that quarter’), and this is reinforced through OSIC’s network meetings with practitioners.

This highlights some of the practical benefits in taking a gradual approach to conformity. It can be better to accept different definitions of data initially to help get the process started, and then align different definitions over time. This can help a gradual change in existing internal reporting processes into something that is comparable across a sector. Where publishing requires some good will and cooperation, lowering the initial barrier to entry is useful in helping the system establish the critical mass that will later make it easier to align on consistent definitions.

Case study: Climate change duty reporting in Scotland and Isle of Man

In Scotland, there is a well developed process for reporting on the compliance of public authorities with climate change duties. The specific areas that authorities have to report on are defined in [Climate Change \(Duties of Public Bodies; Reporting Requirements\) \(Scotland\) Order 2015](#). This has since been [amended by a 2020 order, which added several new questions](#).

This data collection (especially of emissions and emissions reduction project data) is used by [mySociety’s CAPE](#) to make it more explorable and accessible.

This legislation is precise, and contains images of questions as if they were a paper form. Through [an interpretation act](#), this can be adapted into other forms (e.g. an Excel template, or web interface) as long as it does not materially affect the form described. This legislation

requires the form be returned to the Scottish Government. The Scottish Government funds the Sustainable Scotland Network (SSN) to manage reporting, including coordination of reporting, and the receipt and high-level analysis of reports.

Having reporting questions listed in detail in legislation makes sure that all reports work to the same template and questions. However, it has the problem of being restrictive when it comes to updating questions or adding in new areas of reporting. Interpretation and advice on the reporting questions is handled through non-statutory guidance produced and updated by the Sustainable Scotland Network. Additionally, the questions listed in the Scottish Public Bodies Climate Change Duties Reports don't fully address the broad nature of the duties in the Act, including questions relating to the wider influence of public bodies on area-wide emissions (A non-required section on this is currently included in the form). The more fixed the questions are, the harder shifts in question scope and interpretation over time becomes.

An alternative approach highlighted was the more recent example of reporting requirements created in the Isle of Man. The form of the Isle of Man's '[Climate Change \(Public Bodies' Reporting Requirements\) Regulation 2022](#)' leaves much more discretion to the collection process. While noting broad areas that the annual form on compliance with climate change duties may ask about (buildings, power or vehicles used by the authority), it does not detail the exact questions - just saying authorities have to provide ("*such information as may be requested by the Council of Ministers in the annual reporting form in relation to the public body*"). As this collection process will then be run by the Isle of Man government, tweaks to the exact form of these questions, and clarification over meaning will be more simple to administer.

The current situation in Scottish reporting data reflects a challenge between data publication legislation that is detailed and prescriptive, and enabling climate change reporting to evolve over time as data, approaches and standards develop.

If using a voluntary standard, be clear on the benefits to the data publisher

While we think standards need to be required for publishing to work effectively, for many existing standards, it is useful to understand how data publishers can be convinced to publish in more useful ways. The abstract value of joined-up public data is not enough on its own - standards have to be easy to use, and have benefits to the data publisher.

The Local Government Association has created voluntary [schemas](#) for the publication systems required in the Local Government Transparency Code, plus a central repository to store information. However, there are lessons to learn from this voluntary scheme, because it is not well used. The site requires a large number of clicks to get what most data publishers need: a spreadsheet template. Through a lack of requirement to publish in this format and a high technical literacy requirement, the [data repository approach](#) only has a third of councils publishing any data at all, with most councils only publishing a few datasets, and half the datasets published being released by three councils. To unlock the benefits of open data for the public, data production needs to bend to what is practically possible within public authorities.

Where voluntary data coordination schemes work, it is because they are providing clear benefits back to the data publisher. Organisations who adopt the 360Giving data standard for grant information can use 360Giving's data analysis tools to explore that data - in some cases bringing together different grant-making teams in the same organisation for the first time. 360Giving makes it extremely easy to access a template (and almost all data publishers submit in Excel's .xlsx rather than .csv or .json files) and provide support to transition from an existing internal Excel sheet (by providing a spreadsheet that can map between the two formats).

Similarly the near universal compliance with the Office of the Scottish Information Commissioner's collection of FOI processing statistics can be partly explained by their value to people working on FOI. Working with the commissioner's request raises the profile of FOI statistics inside the organisation, and provides benchmarks that can help with internal arguments about good performance.

Case study: Food hygiene ratings

The food hygiene ratings are an example of where a standard is enforced by a regulator without the exact details being specified in legislation.

Food inspections are carried out by local authorities. The Food Standards Agency (FSA) collates information from each council, and then runs a single search page via ratings.food.gov.uk, with the data also available through an API.

In practice, this happens as a result of broad audit and data requesting powers - that have been formalised into an explicit agreement on data transfer.

The Food Standards Agency has powers under the Food Standards Act to issue guidance to local authorities and under the Official Feed and Food Controls (England) Regulations to monitor the performance of local authorities, and request information from local authorities. In practice, this happens through Code of Practice and [the Framework Agreement](#), which specifies the information that authorities need to supply data to the [Food Standards Agency's LAEMS](#) system. These processes are not legally binding, but in practice all local authorities with food enforcement responsibilities comply.

Create a central repository of the location of the published data and aggregate this data

Once data is being published in the same format, the next challenge is to bring all the pieces together. This is a simpler problem to solve than the initial standard, but getting the details right is important to creating robust processes.

Our recommendation is that the role of the data convener is to maintain a register of URLs where the individual data publishers are storing the information. We recommend against requiring the data publishers to publish updated resources at a fixed URL. It needs to be easy for data publishers to inform the convener about the latest version of their dataset, assuming each time it is published the URL changes.

An ideal solution would be an online URL-submission system that validates the data at the submitted URL, and makes the most recently received URL the canonical one. But other solutions using simple CMS forms (or a Google Form or similar) would fulfil the basic requirement that it must be *simple* for data publishers to send in updates.

This is counter to what would be assumed to be good practice by data professionals, but is again, working with the reality of non-technical data publishers working within existing CMS systems. Problems with alternative approaches are explored below.

This submission process is also a good opportunity to introduce data validation tools. Without this step where information is brought back together, it can also be harder to notice where issues with the data standard have emerged in the first place. For emerging datasets or new standards, bringing them together in one place provides key insight into whether the data being produced meets expectations.

It is often a good idea for data conveners to publish an aggregated version of the data to make it easier for data users to obtain all the data at once, and we would recommend this happens where resource allows. That said, for some datasets fast access to the source of the data might be more important than getting all the data at once (see polling case study below). In these cases, aggregating the dataset is less useful than providing an accurate list of source URLs, which can be queried by end users.

Alternative (but not recommended) approaches

Discarded approach: Data publishers have a universal fixed URL system

A potential solution to this across different datasets would be that there should be a standardised technical index on all authority sites that point to the datasets they are required to publish. This was an approach suggested [by Democracy Club in 2016](#). If this system was in place, each new dataset requirement would be issued a new unique identifier, and it would be easy for anyone with a list of authority websites to pull together all the data.

This might be an ideal end state, but given existing systems and difficulties with technical skills, it seems unlikely to be successfully introduced at this time. Intermediate approaches do have value though, and as local authorities often procure from a small number of CMS providers. Encouraging these providers to adopt better handling for dataset re-rerouting could be an effective way of making data more accessible for large groups of local authorities.

Discarded approach: Data publishers have a dataset-specific fixed URL system

In this model, data publishers are always re-uploading the updated information at the same URL, meaning that the data convener just needs to re-query the same list of URLs to get the updated set of data. In principle, this makes the data convener's role simpler - in that less work is required to maintain the register once the set-up work is complete.

This is effectively the position of the Brownfield Land dataset. The problem here (explored in a case study below) is that in practice many of the links are out of date, and more up to date information is available on the original sites, where new datasets have been uploaded without changing the link. The content management systems of authority websites can sometimes make it deliberately hard to update a hosted file to exactly the same URL (to avoid two different 'report.docx' files overriding each other). Given the tools data publishers have (a general purpose CMS), it is much easier to publish a new file to a new URL.

Some authorities have much more ability to handle data publishing as a specialist skill set, while others will be handling data publishing as a side-role of curating the dataset for internal purposes. Requirements to publish have to be sensitive to how the process will be implemented in different areas. This different level of resource makes it difficult for best practice to be required. The practical approach is to make peace with this, and build a register that is easy for data publishers to update, rather than having them update a fixed URL.

Case study: Timely access to polling station data

It is important to remember that fragmentation may not be the only problem with a dataset. For datasets where timeliness and accuracy are important, aggregating can introduce new problems into datasets rather than resolve them.

For polling station data, the best available location to access this information is Democracy Club's wheredoivote.co.uk. This information is sourced from local authorities across the country to provide a single postcode lookup. The [House of Lords report on 'Digital Technology and the Resurrection of Trust'](#) recommended that this should be made easier:

"Local authorities should be required to publish open, machine-readable information on elections, including what elections are taking place, who the candidates are and where polling stations are located."

This would be an improvement on the current situation, but would not address the fragmentation issue. However, Democracy Club staff raised issues that aggregated services can sometimes be helpful to end-users, when time is of the essence. Long-run statistical data might be useful to aggregate, but for other datasets (like polling stations), not having correct information at the time it is needed means the aggregation is not useful. Aggregation vs timeliness is similarly an issue mySociety have encountered when exploring [procurement contracts in local government](#), where aggregate datasets exist, but are not as up to date as to be useful for the service.

The ODI report '[Comparing decentralised data publishing initiatives](#)' argues that where successful, a decentralised approach has the following benefits:

- **Removing central costs for data collection and management**, which will be higher where there is a large number of publishers who may need to contribute, or where the volume of data to be collected is significant;
- **Increasing timeliness of access to data**, for example if the data is regularly updated or published and creating an intermediary would slow down the publishing of that data;
- Making data available from source, rather than indirectly via an intermediary may **reduce risks, increase trust or clarify the provenance of data**.

Sym Roe of Democracy Club made a similar point, that while in principle they wanted to end up with a national dataset of polling stations, an aggregated dataset was not helpful to them if it slowed down the pace of updates or corrections to the data. More aggregation steps move the final dataset further away from the original publisher, which dilutes principles of ownership and

trust in the data. If the publisher of re-aggregated data is not responsible for resolving errors, or does not have timely ways of reflecting upstream updates, it is difficult to trust the accuracy of the information.

As Chris Shaw put it: *“If the problem you set out to solve is fragmentation, the solution you arrive at is an aggregation service of some description. This can work well sometimes, but in some cases it puts a barrier in between the data users and the maintainers of the source information which can be counter-productive”.*

Case study: Brownfield land registry

Brownfield land is previously developed land that is not currently in use. The [Town and Country Planning \(Brownfield Land Register\) Regulations 2017](#) created requirements for local authorities to have a register of this land, the information to include, and that the central government can require this information to be provided “in such form and by such date” as can be specified later. This is used to create a national picture of [where brownfield sites are located](#).

In practice, this is managed through the Department for Levelling Up, Housing & Communities digital land team. They [provide a format and a process](#) for publishing the information. This process involves a template CSV file for listing the required information in the legislation, a validator tool, and instructions on upload.

This process requires that the data publisher upload the file to their own website, and email the URL to the digital land team to have the data imported to the national register. The intention is that the file is a URL that doesn’t change over time, and that future updates will update that same file.

However, the publishing step presents a problem. In reality, many of the URLs submitted are either non-permanent or not updated.

Reviewing the [URLs that have been provided to the Digital Land team](#), many (that have not been specified with an end date) now return page not found errors. In some cases, the filename suggests it is a register for a previous year that has not been updated. For instance, Stockton-on-Tees Council [publishes on their website XLSX files for 2017 to 2021](#), but the central URL registry is only aware of 2017 and 2018 uploads. This means the central registry is missing the six brownfield sites added from 2020 onwards.

This kind of problem is why we recommend working with the limitations of data publishers in local government (who may have access to the CMS, but not the ability to create and update

static URLs), in favour of an approach where it is easier to submit new updates (and can be worked into the process for publishing new data), rather than having requirements for a consistent URL over time, that not all data publishers will be able to guarantee over time.

Ideally, in the long run, improved data management systems in vendor CMS systems should simplify this process on both ends - but in the meantime work needs to be done to work around the limitations of the system. Building processes and technical processes that expect new files rather than updates means manual chasing for new information can be automated.

Provide support and assistance

A theme through our recommendations is that successful data publishing needs more support from the centre to coordinate data publishers and get the best results. As such, we recommend that specific time and resources need to be dedicated by government departments (either internally or contracted out) to get the most value out of the data publication process. While we think a legal framework is required as a backstop to encourage participation and cooperation, the general tone of this function should be supportive, and helping people find ways to make the process work.

This function might include digital tools, such as validation and publication tools, but also administrative maintenance (chasing missed submission dates), working with publishers to maintain the standard over time, and providing support to make sure that processes work for data publishers of varying technical ability.

Importantly this support function has to reflect the existing skills and expertise in the public authorities with the new duty to publish data. As SpendNetwork's Ian Makgill put it “[w]e've increasingly required the parts of government that do 'delivery', e.g. NHS Trusts, to become data publishers, but we've never provided the tools, skills or capabilities to make them good at it”. While we want to make public authorities better data publishers in the long run, in the meantime the support required for each new dataset needs to reflect the reality of the tools and training available to publishers.

Generally it has to be remembered that the staff at public authorities publishing data are generally performing this task as part of other duties. They might have expertise in the subject area, or data publishing, but probably not both, and potentially neither. This means it is important to allocate resources to support, both from any convening organisation and to support peer-to-peer learning (for Scottish climate duties reporting, Sustainable Scotland Network uses a public sector forum to encourage this.). In addition to the technical side of publishing, Tim Davies told us “*much of the journey of supporting adoption of a standard is a community-building and support process*”. This

kind of expertise building is important in making it easier for organisations to effectively comply with requirements to publish.

This extra support comes with a cost that needs to be accounted for, providing support in central staff time and bespoke tools is one of the most effective places for time and effort to be spent. It reduces the costs of the many authorities reporting information, and makes the final dataset more comprehensive, effective and useful. Investment in support is repaid in the value of the final data.

Case study: Technical and non-technical barriers to local government publishing

As part of our research on this report, we talked to [Open Innovations](#), who have been involved in multiple projects to get different public authorities to publish data in a common standard.

These projects found some common issues that need addressing in attempts to improve publishing. In one project involving [business rates from multiple local authorities](#), the approach was to construct a standard that had some features from the different data approaches already in use by the authorities. But even with this half-way step, there was still difficulty in getting people to publish to the standard, because of a lack of tools or specific technical skills available to the team publishing the data. Here, multiple tools were created to help authorities complete and test the data, but it has been a challenge to get it published, and then see it continuously updated correctly to the same format (through changes in internal system, processing steps changed, or different people doing the role).

One factor that cut across the different projects was that the publication of data by local authorities reflected their siloed organisational nature (both internally, and in terms of working across authorities). The pitch of open data is that there is a clear abstract value in everyone publishing the same way (it enables useful tools in the long run) - but this is far from the dominant view of decision makers inside local authorities, where there is a focus on solutions to specific local problems. This also reflects just a difference in resources between authorities. Some authorities have much more ability to handle data publishing as a specialist skillset, while others will be handling data publishing as a side-role of curating the dataset for internal purposes. Requirements to publish have to be sensitive to how the process will be implemented in different areas, and data conveners (even when publication is required) should have an eye on explaining how data publication also helps fulfil local objectives.

Provide validation and publication tools

Something that data conveners can do to help bridge the technical skills required is create interfaces and tools that help data publishers interact with the standard without a highly technical toolset.

Data validation tools can be used to help publishers check if their data meets the standard. These don't have to be heavyweight, and should be appropriate to publishers' skills: they can be built into an Excel template, or a web form; for more advanced publishers they could be command-line or API-based validators. 360Giving provides a website that can give feedback that the submitted spreadsheet conforms with the standard, and highlights errors to fix (and does some basic checking around wrongly formatted columns).

This kind of error checking should be well integrated and signposted to non automated forms of support. As technologist Peter Wells told us, *“[t]he tone of these [platforms] and the support can be vital. It should be seen as a support function as much as - or sometimes more than - an enforcement function. Tools like this can offer automated tips on compliance, or simply provide a route to an - often better resourced - national data team that can help”*. As the goal is to bridge the gap with non-technical users, error messages should take care to be understandable.

An alternative (and sometimes overlapping) approach is **publication tools** to help organisations upload or publish their data in standard form. These can be minimal, like an Excel template, or a simple web form, or more advanced like endpoints to upload JSON, depending on publishers' skills and resources.

For instance, to create a dashboard of diversity information for organisations in the Leeds City Region, Open Innovations [created an online interface](#) where organisations could complete an audit for their organisation (or parts of their organisation), and download the final data as a CSV to submit. This approach has the advantage of being able to include detailed descriptions for each field, and live checks of whether an entered value is valid.

For tabular data (a grid of rows and columns), a publication tool may validate the data passing the standard before submitting the link to the resource to be stored in the database. For data gathering processes that are not tabular data but contain lots of different structured information or questions, online forms are a good way of expressing the requirement in a way that provides quick feedback of problems, while minimising the ability for users to make mistakes with an Excel template (see the case study below).

In the past, external software suppliers to public authorities have sometimes charged fees for extracting data from databases, making it harder for authorities to publish data. In line with the

Government's procurement guidelines in the [Digital and Data Playbook](#), contracts should be drafted to ensure that authorities can get access to their own data without financial penalties.

Case study: Investment in good tools means the data convener's job is easier in the long run

More complex data gathering needs effective technical approaches. Work put into data collection and formatting tools upfront save time and energy down the road.

Bespoke tools have benefits over adapting common survey platforms. The Office of the Scottish Information Commissioner (OSIC) collects information on FOI statistics from Scottish public authorities online. When the OSIC moved from a custom web app to a generic form provider, the time taken to administer the process increased. Good investment in a data gathering process can reduce the need for as much support for data publishers.

For example, there can be problems when authorities use Excel sheets to collect free text information as well as tabular data. We would advise caution in using Excel to store significant amounts of free text beyond a prototyping process.

From the experience of Sustainable Scotland Network, who coordinate the returns of Scottish climate change duty data, it creates a number of challenges, both in terms of user-friendliness of data reporting and post-reporting analysis and data visualisation. This leads to time-consuming technical queries that would be resolved by a better data collection system (such as an online form with input validation), as well as challenges in collating and handling the data once reports have been published.

Spreadsheets can seem like a quick method of collecting information, but in the long run investing in a bespoke online form would make the outputs more usable and comparable between organisations. Time spent running technical support on a spreadsheet-based form is better spent engaging with more substantial questions.

Conclusion

Fragmented data is a legitimately hard problem. Solving it requires using technical and non-technical skills to improve coordination and cooperation between people working in many different authorities. But the rewards of solving it are very large - bringing together data from across local government or the NHS means we can make better decisions and better understand the impact of previous decisions.

Our recommendations are responses to common problems. Not all solutions are suitable in all circumstances, but the general principles can be adapted to fit new situations.

Our key point is that more work upfront by the data convener is less work for data publishers and data users, and this leads to a dataset that can have a much bigger impact. The lack of this investment at the centre means that the promised returns of transparency may never arrive.

It is also important to remember that processes that work need to function within the skills and limitations of data publishers. While the long term goal of data publishing in government should be to have more skills available in more organisations, data processes at the moment need to work for the least technical link in the chain to gather information from all authorities.

Fragmented data is a hard problem, but it is one that can be progressively chipped away at, and we hope our recommendations can be an important step in making public data more accessible and useful.

Acknowledgements

This report was informed by conversations with a range of organisations. Some additional comments made through the feedback survey have fed into our work, even where we have not used direct quotes. Many thanks to everyone who contributed their time and experience.

- 360Giving - Katherine Duerden
- Climate Emergency UK
- Connected by Data - Tim Davies
- Democracy Club
- Sustainable Scotland Network
- Spend Network - Ian Makgill
- Open Innovations - Stuart Lowe
- Office for Scottish Information Commissioner - Claire Stephen, Paul Mutch
- Peter Wells
- Matthew Waddington