Statistical Data Publishing Workshop

Report V1.1

Author: Mark Braggins

Executive summary

This report summarises discussions during the Statistical Data Publishing Workshop. It documents existing tools, identifies 'pain points' or gaps in the publishing process. Lastly, it suggests potential tools which could help improve statistical data publishing.

Background / purpose

The Southern Policy Centre was established in 2014 as the think tank for central southern England. It specialises in improving public policy making, and conducts research into the social challenges facing society, including devolution, poverty and exclusion, ageing population, and social exclusion. Southern Policy Centre uses open data for its research projects, and hosts ODI Hampshire, part of the Open Data Institute's worldwide network.

On 27th September 2017 Southern Policy Centre (SPC) ran a half day workshop entitled ODI Hampshire: Statistical Data Publishing workshop. The venue was the Conference Centre at Southampton Solent University. The workshop was commissioned by the Open Data Institute (ODI), as part of a programme of work around tools for publishing open data. The workshop was organised using Eventbrite, and there's a brief introductory blogpost on the SPC blog.

The purpose of the workshop was to:

- list the tools being used by the publishers of statistical data
- discuss what those tools are being used for
- identify gaps or 'pain points' in the publication process
- suggest potential solutions for some of these problems, and recommend next steps.

The exact format of the workshop was left to SPC to decide, but the ODI provided three core hypotheses as a guide to what they were looking for:

1. **Quality** - The quality of open data being published today and the quality of how this data is being published are often too low to enable effective discovery or reuse. Our hypothesis is that better tools (for e.g. cleaning data, describing it or simply better integration of the various steps of the publishing process) may improve this situation.
2. **Speed** - Publishing quality open data takes too long. Our hypothesis is that it is possible to improve tools and methodologies for open data publishing that decreases the time it takes someone to publish open data.
3. **Cost and automation** - Publishing quality open data today is a costly process involving a large ratio of manual, tedious processing. Our hypothesis is that there are opportunities to automate the process of publishing open data, increasing

speed, quality and reducing cost.

This report is a summary of discussions

Agenda and participants

The agenda for the morning was:

1. Welcome
2. Introductions: you, your organisation, the kind of data you publish
3. Which tools do you use?
4. What do you use them for?
5. What gaps or 'pain points' have you experienced?
6. Do we need better tools?
7. If so, what would they look like?
8. Recommendations / Identify next steps

The intended audience was organisations who publish statistical data, or who provide tools for data publishers, but anyone with an interest in publishing data was also welcome to attend.

There were 17 attendees in total, from a range of organisations including:

- Barter for Things
- Department for Communities and Local Government (DCLG)
- Drawnalism
- Famiio
- Geodata Institute
- Isle of Wight Council
- Office for National Statistics (ONS)
- Open Data Institute (ODI)
- Ordnance Survey (OS)
- Southampton City Council (SCC)
- Southern Policy Centre
- Swirrl
- University of Southampton (UoS)

Attendees introduced themselves and their organisations at the beginning of the workshop.

Tools

To begin with, participants were asked to spend a few minutes listing tools they use as part of the data publishing process. There were 93 contributions in total, which participants wrote on

post-it notes. See Appendix 1 for the full list.

Similar tools were grouped together. The groupings which emerged were:

Sites, portals and paper

The first category identified those tools which publishers use to 'get the data out there'. These included websites like ONS and Ordnance Survey, publishing platforms like PublishMyData, and portals like the London Datastore and the European Data Portal. The ONS uses additional tools in conjunction with its website, such as DataBaker (a bespoke content management system, built by ONS).

There are many tools designed specifically for publishing, and other tools which can be used for publishing, like Google Sheets and Sharepoint. Cloud services were also mentioned, like Google Cloud and AWS. The group didn't identify an exhaustive list, and the workshop had access to the ODI Tools audit for reference.

Dissemination and communication

Dissemination and communication covers any tools which publishers use to convey information to the effect that there is some data available. This includes social media tools such as Twitter and Facebook, and also blogs and email newsletters. Paper was also mentioned as a low tech tool useful to reach some audiences.

Spreadsheets

Spreadsheets were suggested by many participants, particularly Excel and Google Sheets, non-proprietary tabular data such as CSV, and open source equivalents like LibreOffice. Web forms and MarkLogic forms were also identified as being useful for data collection.

Conversion

Another category was for those tools which enable data to be converted from one state to another. Some tools, like Grafter (a software library for ETL), and Received by SQL (a converter to Excel) were created in-house. Others, like DataBaker (for transforming XLS to structured CSV), and Florence (Bespoke CMS for ONS) are open source. Others were specific to certain databases, such as SPSS, SQL Developer, and Toad for Oracle.

Analysis

Are protocols and standards tools? Well, participants certainly thought so, listing OGC standards for both data and delivery as being useful for analysis. R was identified for checking and analysing data, and Excel, SQL Server, SPSS and Microsoft Access were all listed as tools for data storage and analysis.

Management

Management tools are widely used, with 18 tools identified by the group. These ranged from open source databases used for data management, such as: MySQL, PostgreSQL, PostGIS and Sqlite, to portal / management development tools such as PHP and Python.

Spatial

Spatial tools were mentioned many times, i.e. tools that focus on place or location. A large proportion of statistical data relates to places. The collection of places is large, complex, and is updated quite frequently. It is also an important term of reference information. Many spatial tools also fall within different categories, such as analysis and visualisation, but the group felt spatial worth identifying as a category of its own.

Visualisation

This was the category with the largest number of tools. These include tools for mapping of spatial data, such as: OpenLayers, Leaflet, QGIS, PostGIS, ESRI, Google Maps, Cesium and OpenStreetMap, and tools for visualising statistics such as R+, Shiny, D3 (and other JS libraries). Minecraft was mentioned as a way to visualise geospatial data in an entertaining way, but also as a way to help discover new datasets that may not otherwise have been spotted.

‘Pain points’ and gaps

For the next main section of the agenda, participants were asked to spend a few minutes identifying problems and ‘pain points’ they associate with publishing data. As with tools, attendees wrote on post it notes, which were put on the wall and grouped together. See Appendix 2 for a full list. All of the groupings identified during the workshop fit within at least one of the ODI’s three core hypotheses of Quality, Speed, Cost and Automation.

This time, participants were also asked to identify what in their experience is the most challenging problem or gap. This prompted discussions around half a dozen themes:

- Quality

Participants identified many issues associated with one of the ODI’s core hypotheses, Quality. One of these relates to trust in data quality. In order to trust data, users need confidence in its source, and the method of collecting and publishing the data. It is also important to understand what is missing. As one participant put it:

> “some data quality is very good because the pathway to capturing that data and the system is very good. But, if we were to look at that without understanding the other complex scenarios in which data is poor, you could start drawing false conclusions.

> And it's really important to understand what you are not capturing, why, and what that might tell you. So, understanding the gaps."

Other factors include the reliability of the publishing platform, dirty and inconsistent data received from other sources, lack of adoption of data standards, and lack of automation for quality scoring. Some datasets can be extremely large, resulting in lengthy processing time.

- ## Standards

There are many pain points and gaps associated with standards, or the lack of standards. Common issues include:

- Lack of a common format
- Insufficient consistency in data presentations (no two Excel files look alike)
- Lack of (good quality) metadata to support statistical data
- Absence of standards
- Where there are standards, they aren't always adhered to

> "...you've got the hospital episodes dataset, and you've got the deprivation dataset, and you're trying to work out how they relate to each other. So, are they describing places in the same way as each other or, if they're not, how are these things connected?"

- ## Discovery

Another pain for publishers is how best to ensure that your data can be found and used by others. Making data findable takes time and effort, and there is no single right way of doing it.

To paraphrase one of the workshop attendees:

> As a user, you may have a big and complex problem that you are trying to gather evidence on, and to do that you need lots of data from lots of different places of varying quality.
>
> Some datasets are really really big, and you only want a bit of it, but also there's thousands of datasets to choose from, and you might be missing stuff that's useful, or using the wrong things...
>
> If we could make our user's problem easy, then we would have published well.

- ## Ethics

Some organisations, such as local councils, hold detailed data, which they use to deliver services, but cannot - and would not want to - publish at a level where individuals could be

identified. Publishers always need to consider ethical issues before publishing data, particularly at lower levels of geography.

> “So, you’ve got limited funding, resources and you could be really well-informed in order to target your services - making the most of those resources but what you now know could be intruding on people’s privacy. Using that information to do your job is one thing, but publishing it is a different matter.”

> “If you want to apply evidence to a policy decision you should be able to show your working, to say this is the evidence I used, and the assumptions I made, and so all of that evidence is available to other people and you can link to it, or make it available sensibly, then somebody can question your conclusion.”

On a positive note, Government statisticians have a very well defined set of guidelines on disclosure control which helps publishers determine what statistics are suitable for publication

- ## Audiences

Publishers of statistical data need to spend time making complex data easy for users to work with, and finding ways to reduce the likelihood of users drawing false conclusions from the data.

Some audiences care a lot about quality, and others less so. Some care about currency and are happy to be signposted to a feed. Others might be completely non-technical, and non-expert: “I just want to know ‘the number’ for where I live and compare it to where my auntie lives”.

So, publishers need to be cautious of trying to come up with a single solution, and can’t possibly satisfy every conceivable need. The consensus at the workshop was that publishers of statistical data should try to accommodate specialist, expert users, and also provide something for the less technical, casual users.

A popular suggestion was for publishers to make friends with friendly local storytellers within the organisation - “because you need somebody who can talk to different audiences”.

- ## Data management

Some organisations - like ONS and the Geodata Institute - encounter issues collecting data from different sources. They receive data in a variety of formats, which have been sent to them by other organisations using different labels and descriptions. There is usually a need to cleanse the data, and the lack of common formats and structures really doesn’t help the publishing process.

> “Good data management is like getting your house in order, because the data starts off closed, it doesn’t start open. The more that can be done upstream, the better.”

> “It’s always different depending on who is providing the data, and that takes a massive amount of resource, in the end to cleanse, to harmonise, to standardise, before we can start to worry about other things. If that could be resolved, then the flow of the data through the systems would be much, much quicker.”

Other common problems include lack of resource, difficulties associated with legacy systems, cost, and the amount of effort required to protect and mask data.

Organisational culture

Another area that participants highlighted is organisational culture, such as fear of criticism:

A common issue in many organisations is that data collected for one purpose subsequently becomes a candidate for publication as open data. Some data owners fear the audience finding out that the data quality is poor, and they worry about a potential backlash. This is quite common in local government, particularly for those councils who lack constructive relationships with local communities and journalists. Even where there is senior executive support, institutional resistance to opening up existing datasets can be very strong, particularly if the data owner fears criticism as a result of what the data shows.

> “Institutional resistance - we had a lots of pushback from people who basically know that their database is s**t - it’s fine for their purposes, but they never wanted it opened and other people to see it.”

For example, most councils have pre-defined routes for gritters to follow in the event of bad weather. Routes might have originally been drawn on paper maps, and then transposed - probably by junior staff - onto their council’s Geographical Information System (GIS). When the data was originally captured, accuracy might not have been important - staff and contractors just needed to be able to tell which roads had gritting priority, or weren’t going to be gritted at all.

Many councils show gritting routes as maps on their websites, and there are good reasons why it is in the public interest for councils to publish the underlying data so it can be reused by others. However, some councils have been reluctant to publish the data. One possible explanation might be that - when inspected closely e.g. using a Geographical Information System (GIS), routes may appear to run through people’s gardens. This data was never intended to be visible to the public even though it should be open data.

Do we need better tools? If so, what would they look like?

Having established what tools are already being used, and what ‘pain points’ and gaps exist, participants were asked to split into three groups. Keeping in mind the three core hypotheses of

Quality, Speed and Cost and Automation, each group was asked to suggest three tools that could make a real difference for data publishers, and address some of the pain points identified earlier in the workshop. This was part wish-list, and part action plan, also referred to as 'Pain Reliefs':

1. Data conversion utility

Tool to convert data to a desired state.

The example discussed was data collected by councils using a legacy system which has no mechanism to check postcodes. The tool would accept data from the legacy system, and convert it to the desired state, including checking and converting postcodes into a standard format. Similar issues exist across all councils, so the tool has potential to be used across the whole sector (around 142 councils).

Benefits: Reduces manual effort, pain, and helps maintain common standards.
Hypotheses fit: Quality, Speed, Cost & Automation

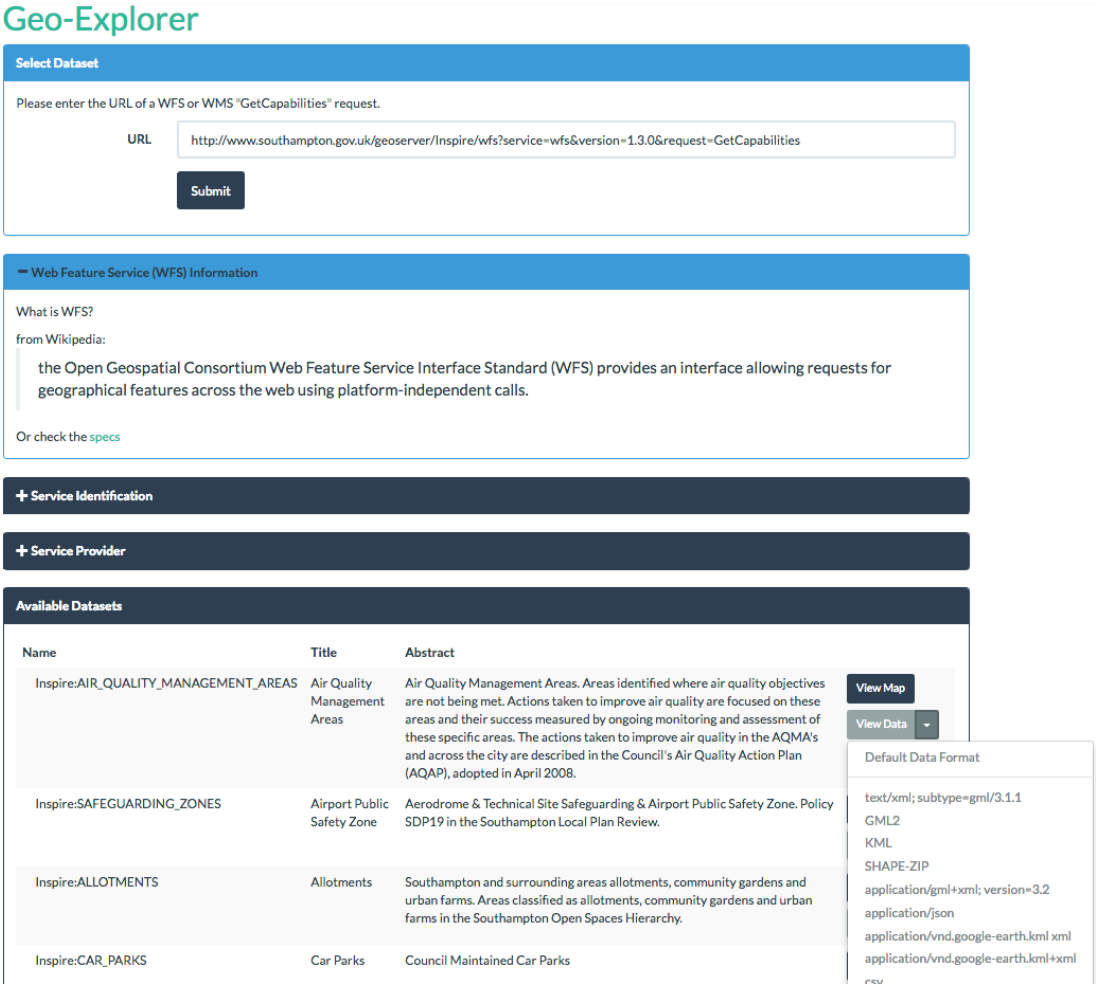
2. Tool to convert jargon format

Some data are published in ways that the majority of people are unfamiliar with, such as Web Map Service (WMS) and Web Feature Service (WFS). While this suits specialist users, it excludes a much larger potential audience.

This tool is to help get data from a narrow, expert knowledge area into a form where people can say - "give me that data, so I can use it in my spreadsheet"

A version of this tool does already exist for geospatial data, and is known as Geo-Explorer, created as a mini-project at the 2016 Research Festival of the Web and Internet Science Research Group at the University of Southampton. The code is available on Github.

Benefits: Reduces pain, and manual effort, and makes data available to a wider audience.
Hypotheses fit: Quality



3. Tools for data collection

The third tool helps others collect data in a form which is as readily usable as possible.

“I want to report a thing: pothole, broken stile, fallen lamppost, or whatever, and I’ve got a smartphone, with my app of choice. I want to convey this thing that I’ve discovered to the organisation which is responsible for dealing with it.”

Capturing precise, good quality data in a structured way is a major step towards creating an open data ‘pipeline’ or ‘manufacturing process’.

The ‘tool’ in this case could be the wide adoption of an existing standard such as Open311:

> “Open311 is standardised way for computers to report problems (like potholes or fallen trees) to the computers run by the bodies that can fix them (like local governments or city departments).”[1]

[1] https://www.mysociety.org/2013/01/10/open311-introduced/

Collecting the data in a standardised way simplifies the process of producing statistics based on that data.

Benefits: Helps improve quality, reduces pain, prepares the way for automation, and helps maintain a common standard.
Hypotheses fit: Quality, Speed, Cost & Automation

4. Error checking

The next tool would check for errors in the supply of data. This could include checking that data has been transposed correctly, validating and reporting back any errors so that data comes back in a more consistent, harmonised way.

Benefits: Helps automation, thereby reducing manual effort and pain
Hypotheses fit: Quality, Speed, Cost & Automation

5. Tool to add metadata to datasets

Metadata lets users understand what data a dataset contains, how it was collected, by whom etc, and make a sound judgement about what that data can be used for. Good quality metadata helps to reduce the chance of misuse of that data going forward. However, adding metadata can be time-consuming and tedious for the publishing organisation. This tool would assist organisations to add metadata to datasets, minimising the amount of manual effort involved.

Benefits: Helps reduce pain, free up staff time, improve speed.
Hypotheses fit: Quality, Speed, Cost & Automation

6. Tool to aid discoverability

The next tool would assist publishers in making their data more discoverable. This might include a way to support more precise searches for data, both so that users get back what they were looking for, but could also know when it doesn't exist. So, tell the difference between "I haven't found it" and "I know it's not there".

One potential method might be to

> "have a strict taxonomy for data which publishers could use when describing datasets, so that users could 'browse the spectrum'. With statistical data there could be ways to use that known structure of the data to say 'what are the measures and units of these datasets, do these come from some kind of controlled list', which again comes back to standards, which is always tricky, but at least within the sphere of government publishing of statistics.

> With online statistical data in a known structure with some machine readability, you could answer that question precisely potentially, because you can either stick it in metadata or you can look into the data and see what are all the areas this dataset is about - are any of the wards in Southampton, say. So, the ability to support more precise searches could help discoverability, and that seems feasible in the world of statistics, if not easy.“

Benefits: Reduces pain and manual effort, and opens up data to a wider audience.
Hypotheses fit: Quality, Speed, Cost & Automation

7. Tool to clarify requests

The next tool would help organisations respond to management or FoI requests. This would assist the organisation to say how it has interpreted the request or question, together with any clarification and context. The tool would process the language used in a request, and respond:

> “This is how I have interpreted your request, do you agree or can you clarify that? This is my response to what I think you are asking, and this is all of the context with which I wish the answer to be placed.”

Benefits: Helps improve quality, reduce pain and manual effort.
Hypotheses fit: Quality, Speed, Cost & Automation

8. FoI common response

Public authorities - and many organisations delivering services on their behalf - are subject to the Freedom of Information Act. The main principle behind freedom of information legislation is that people have a right to know about the activities of public authorities, unless there is a good reason for them not to. As a result, many thousands of requests are received each year, across central and local government.

FoI requests are generally for information that is disclosable, and can be a very expensive. A tool that takes the most common kind of FoI request and automates them for the general public could be very powerful in terms of just reducing cost to the public purse and making information more available to people. The current process is very piecemeal and inconsistent across different bodies. Some common standards for FoI requests could both save money and also help disseminate information more effectively.

There is scope for a single tool, or several. For example, a tool to analyse incoming FoI requests to identify data or services that should be prioritised. Building something that releases data relating to a particularly labour intensive FoI in terms of volume of requests is going to provide value for money, because the cost of servicing a large body of particular FoI requests will be removed.

Having a tool to analyse the data and discover patterns would be very useful, and would enable central and local government to make better decisions about what to do with this sort of data.

Another tool - outside of FoI, but along similar lines - concerns written and oral questions received through the national democratic process, whereby MPs might ask a specific department e.g. “tell me how many school leavers went into tertiary education in the last decade”. These requests require civil servants to research the question, prepare an answer, and then for a minister to write back. The tool could provide an automated way of handling questions and packaging answers automatically.

> “All statistical data falls the same way...you have measures and dimensions, and those dimensions can have a finite set of values generally, and those stats have a data model that is widely adopted and syntaxes for doing it.
>
> The guidance for a publisher is you try and follow those standards. Statistics are a low hanging fruit because, although it’s wide-ranging and complicated, it all follows the same sort of patterns.”
>
> “If there was a way of drilling down with a facility that allowed you to find out things right across 52 local authorities, without any staff interaction, or overhead, that would be incredibly valuable. The reality is that these are burdensome tasks and quite often they are burdensome of staff who don’t usually have access to that information, and have to go out of their way, because they are tasked by the FoI officer. Usually it boils down to a middle manager who hasn’t got the time, and it isn’t really their job, and it may not even be their data, especially if it’s financial for instance. So, I think it would be a coup.”

Benefits: Helps reduce cost and pain, increase automation
Hypotheses fit: Quality, Speed, Cost & Automation

9. Data mask and suppression

The final tool would help publishers by making data ready for safe public use.

Micro-level data would be held in a repository and the tool would automatically mask and suppress data if certain conditions were met. This would safeguard the data and help reduce risk aversion to data publishing.
Benefits: helps reduce risk, increase confidence
Hypotheses fit: Quality, Speed, Cost & Automation

Summary and next steps

This half day workshop identified nine tools with the potential to help tackle some of the issues associated with publishing statistical data. One of those tools is known to already exist, but is

not widely known.

The three core hypotheses: Quality, Speed, and Cost and automation were not stipulated as groupings in advance, though participants were already aware of their existence. Participants worked together to identify the groupings for both tools and 'pain points' or gaps, and what emerged during the workshop fits within at least one of the three hypotheses. Possibly the least obvious fit is Ethics - should I publish? But ethical considerations take time, and the amount of time can be minimised if standards exist and are being adhered to. Once the data is held in a standardised way, publication becomes easier to automate, thereby reducing costs.

Given that this was a half day workshop, there was limited time to explore any of these ideas in detail[2]. The next steps would be to look into each of these suggestions in greater depth, focusing on those which appear to provide the greatest impact.

Freedom of Information was identified as a priority theme, worth exploring in more detail. There are potential cost savings, speed improvements and opportunities for automation across central and local government, and the same tools could be useful across a wide range of organisations.

Appendices

Appendix 1 - Text from post-it notes: Tools

View this table in Google Sheets

Topic	Category	Post-it comment	Link (added later)
Tools	Conversion	Data Baker - Transforming spreadsheets transforming XLS to structured CSV. Open source software designed by Sensible Code Company for ONS	https://github.com/sensiblecodeio/databaker
Tools	Sites, portals and paper	Florence - Statistical data publishing mainly spreadsheets. Bespoke CMS built by ONS for it's website	https://digitalblog.ons.gov.uk/2017/03/03/what-weve-learnt-about-building-a-publishing-platform/
Tools	Sites, portals and paper	PublishMyData: Swirrl platform - ONS Geography, stats publication at DCLG, Scottish Gov, NHS	http://www.swirrl.com/

[2] Some of these issues have been further discussed at Open Data Camp, in a session entitled 'Getting more value from Open Data publishing', led by Christopher Gutteridge. See collaborative Google document: https://is.gd/m0MH3C

Tools	Sites, portals and paper	ONS website	https://www.ons.gov.uk/
Tools	Sites, portals and paper	EU Open Data Portal	https://data.europa.eu/euodp/en/home
Tools	Sites, portals and paper	NOMIS website	https://www.nomisweb.co.uk/
Tools	Sites, portals and paper	Eprints Repository tool	http://www.eprints.org/uk/index.php/eprints-software/
Tools	Management	Data Management. Open source databases: MySQL	https://www.mysql.com/
Tools	Management	Data Management. Open source databases: PostgreSQL	https://www.postgresql.org/
Tools	Management	Data Management. Open source databases: PostGIS	http://postgis.net/
Tools	Management	Data Management. Open source databases: Sqlite	https://www.sqlite.org/
Tools	Management	Data Management. Open source databases:SpatiaLite	https://www.gaia-gis.it/fossil/libspatialite/index
Tools	Sites, portals and paper	Global Value Exchange	https://www.gaia-gis.it/fossil/libspatialite/index
Tools	Sites, portals and paper	OS Open data	https://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html
Tools	Sites, portals and paper	PublishMyData - disseminating on web; managing data collection	http://www.swirrl.com/
Tools	Sites, portals and paper	ONS website	https://www.ons.gov.uk/
Tools	Sites, portals and paper	London datastore	https://data.london.gov.uk/
Tools	Sites, portals and paper	PublishMyData	http://www.swirrl.com/
Tools	Sites, portals and paper	ONS Geographic Portal	http://geoportal.statistics.gov.uk/
Tools	Analysis	Data protocols: OGC standards for data	http://www.opengeospatial.org/standards
Tools	Analysis	Data protocols: OGC protocols for delivery - WMS, WFS etc	http://www.opengeospatial.org/docs/is
Tools	Analysis	DFE: XML and collect	

Tools	Conversion	In house: Received by SQL; Converter to Excel	
Tools	Conversion	Data Baker - transforming XLS to structured CSV	https://github.com/sensiblecodeio/databaker
Tools	Conversion	Grafter (our own) software library for ETL - especially when creating RDF as output	https://github.com/Swirrl/grafter
Tools	Conversion	Grinder - table to RDF tool. Home made open source	https://github.com/cgutteridge/Grinder
Tools	Management	Hedgehog dataset publication workflow automation. Home made open source	
Tools	Management	Open source portal / management dev tools - PHP	http://php.net/manual/en/intro-whatis.php
Tools	Management	Open source portal / management dev tools - Python	https://www.python.org/
Tools	Management	Open source portal / management dev tools Node GoLang etc	https://github.com/DakshMiglani/node-golang
Tools	Management	Open source portal / management dev tools - Web tools and standards (W3C)	https://www.w3.org/standards/
Tools	Conversion	GDAL - converting / processing geo data	http://www.gdal.org/
Tools	Spreadsheets	Excel spreadsheets	https://products.office.com/en-gb/excel
Tools	Spreadsheets	Google Sheets	https://www.google.com/sheets/about/
Tools	Spreadsheets	Excel spreadsheets	https://products.office.com/en-gb/excel
Tools	Spreadsheets	Data collection: Excel forms	https://support.office.com/en-gb/article/Simplify-data-entry-with-a-data-form-00bfc75e-f675-46ad-8703-291fc03e4d77
Tools	Spreadsheets	Data collection: Web forms	
Tools	Spreadsheets	Data collection: MarkLogic forms	https://github.com/marklogic-community/exampleforms
Tools	Spreadsheets	Spreadsheets LibreOffice	
Tools	Spreadsheets	Spreadsheets	
Tools	Spreadsheets	Partner Agencies: Excel	https://products.office.com/en-gb/excel

Tools	Sites, portals and paper	Sharepoint	https://products.office.com/en-us/sharepoint/collaboration
Tools	Spreadsheets	MS Excel	https://products.office.com/en-gb/excel
Tools	Sites, portals and paper	Sharepoint	https://products.office.com/en-us/sharepoint/collaboration
Tools	Sites, portals and paper	Sharepoint	https://products.office.com/en-us/sharepoint/collaboration
Tools	Management	Validation: does the data match the agreed structure	
Tools	Management	Data publication: PublishMyData (Stardog)	http://www.stardog.com/
Tools	Management	Data publication: Elastic Search	https://www.elastic.co/products/elasticsearch
Tools	Management	Data publication: Web page;Excel, CSV, PDF, Word doc	
Tools	Management	Graph database (Stardog) - storing and querying data	http://www.stardog.com/
Tools	Management	My PHP	http://php.net/my.php
Tools	Management	SAP Business Objects - web intelligence, dashboards, SAP Lumira	https://www.sap.com/uk/products/bi-platform.html
Tools	Management	SAP Lumira	https://saplumira.com/
Tools	Conversion	SPSS	https://www.ibm.com/analytics/us/en/technology/spss/
Tools	Conversion	Post Gre SQL	https://www.postgresql.org/
Tools	Conversion	Toad for Oracle	https://www.quest.com/products/toad-for-oracle/
Tools	Conversion	SQL Developer	http://www.oracle.com/technetwork/developer-tools/sql-developer/what-is-sqldev-093866.html
Tools	Conversion	Florence Bespoke CMS for ONS	https://github.com/ONSdigital/florence
Tools	Analysis	Data storage and analysis; Excel	https://products.office.com/en-gb/excel
Tools	Analysis	Data storage and analysis; SQL Server	https://www.microsoft.com/en-us/sql-server/sql-server-2016
Tools	Analysis	Data storage and analysis; Access	https://products.office.com/en-gb/access
Tools	Analysis	Data storage and analysis; SPSS	https://www.ibm.com/analytics/us/en/technology/spss/

Tools	Visualisation	Visualisation and mapping: OpenLayers	https://openlayers.org/
Tools	Visualisation	Visualisation and mapping: Leaflet	http://leafletjs.com/
Tools	Visualisation	Visualisation and mapping: OSM	https://www.openstreetmap.org/#map=6/54.910/-3.432
Tools	Visualisation	Visualisation and mapping: Cesium	https://cesiumjs.org/
Tools	Visualisation	Visualisation and mapping: Stats: R+	https://www.r-project.org/
Tools	Visualisation	Visualisation and mapping: Shiny	https://shiny.rstudio.com/gallery/
Tools	Visualisation	Visualisation and mapping: D3	https://d3js.org/
Tools	Visualisation	Stats: R+	https://www.r-project.org/
Tools	Visualisation	Stats: Shiny	https://shiny.rstudio.com/
Tools	Visualisation	Stats: D3	https://d3js.org/
Tools	Sites, portals and paper	Social media channels: Twitter	https://twitter.com/
Tools	Sites, portals and paper	Social media channels: Facebook	https://www.facebook.com
Tools	Sites, portals and paper	Social media channels: blogs	
Tools	Visualisation	Minecraft	https://minecraft.net/en-us/?ref=m
Tools	Visualisation	QGIS	http://www.qgis.org/en/site/
Tools	Visualisation	D3 (and other JS libraries) Viz	https://d3js.org/
Tools	Visualisation	Google Maps - simple geographical visualisation	https://www.google.co.uk/maps/
Tools	Visualisation	OpenStreetMap - simple geographical visualisation	https://www.openstreetmap.org/#map=6/54.910/-3.432
Tools	Analysis	R - checking and analysing data	https://www.r-project.org/
Tools	Visualisation	High charts	https://www.highcharts.com/
Tools	Visualisation	ESRI	https://www.esri.com/en-us/home
Tools	Visualisation	PostGIS	http://postgis.net/
Tools	Visualisation	PostGRES	https://www.postgresql.org/
Tools	Visualisation	Geoserver	http://geoserver.org/

Tools	Visualisation	R Shiny - visualising data	https://shiny.rstudio.com/
Tools	Visualisation	ESRI	https://www.esri.com/en-us/home
Tools	Visualisation	D3 for data visualisations	https://d3js.org/
Tools	Sites, portals and paper	Cloud servers Google Cloud running online services	https://cloud.google.com/
Tools	Sites, portals and paper	Cloud servers AWS running online services	https://aws.amazon.com/
Tools	Visualisation	Leaflet.js library for visualising map and Geo data	http://leafletjs.com/
Pains	Conversion	Graphite PHP, RDF Lib; homemade; Open source	

Appendix 2 - Text from post-it notes: Pains and gaps

View this table in Google Sheets

Topic	Category	Post-it comment
Pains	Quality	Trusting data: source; meaning; method
Pains	Quality	Reliability of platform
Pains	Quality	Data currency ("Old data")
Pains	Quality	Trust in data quality
Pains	Quality	Variable data quality
Pains	Quality	Poor data quality
Pains	Quality	Unusable spreadsheets ('Raw' CSV files)
Pains	Quality	Dirty / inconsistent data from providers
Pains	Quality	Lack of data standards (+adoption)
Pains	Quality	Data quality
Pains	Quality	lack of 'live' data
Pains	Quality	Data quality
Pains	Quality	Very large datasets take time to process
Pains	Quality	Quality Scoring (Lack of automation)
Pains	Quality	Data Currency: being up to date; knowing what it is
Pains	Standards	Lack of standardisation used in data releases (code lists, classification, registers)
Pains	Standards	Lack of a common format (standardisation)
Pains	Standards	Time taken to configure from SQL to Excel
Pains	Standards	Problem: manual processes
Pains	Standards	Lack of consistency in data presentations (no one XLS looks like another)
Pains	Standards	Lack of unique identifiers - too many IDs for same thing
Pains	Standards	versioning - stats get improved / corrected. Sometimes need access to old versions
Pains	Standards	Legacy systems driving bad habits
Pains	Standards	No common formats used (XLS, CSV, ODS, RDF, SDMX)
Pains	Standards	Standardisation - agreeing with others how to

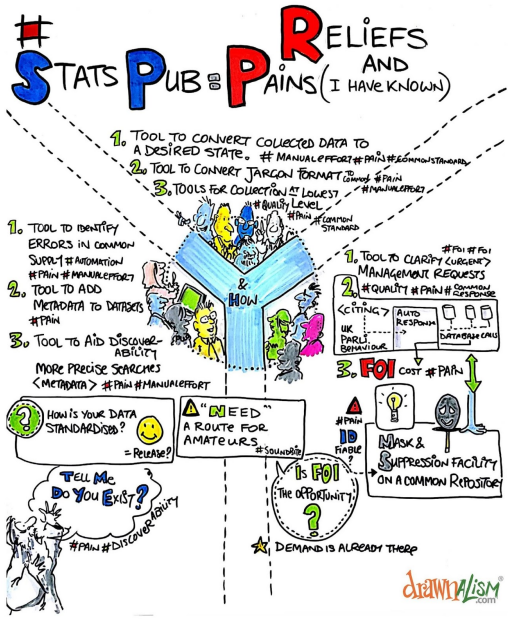
		represent data
Pains	Standards	Interpretation of the question - could be answered many ways but lack of the requestor knowing what they want
Pains	Standards	Interoperability
Pains	Standards	Lack of metadata to support statistical data
Pains	Discovery	Discoverability
Pains	Discovery	Missing data - what's not shown may be more valuable
Pains	Discovery	Making datasets findable
Pains	Discovery	Finding data - knowing where to stop
Pains	Discovery	Issue: no integration standards for process
Pains	Discovery	Effort needed to clean data
Pains	Discovery	Lack of understanding of what open data is & what is open data
Pains	Standards	Gaps: lack of standards for collection
Pains	Discovery	Comparing apples to pears
Pains	Discovery	Many manual processes to match and map data before release
Pains	Standards	Standards not adhered or changed
Pains	Ethics	Data owners recycle IDs
Pains	Ethics	Consistency
Pains	Ethics	Data dump versus API
Pains	Ethics	Different users want different formats CSV / RDF / JSON
Pains	Audiences	Making complex data easy for users to work with
Pains	Audiences	Drawing false conclusions
Pains	Audiences	Complexity of admin and census geography (and that it changes over time)
Pains	Audiences	Loss of data context
Pains	Audiences	Data owner fears others seeing their data is crap
Pains	Data management	GDPR
Pains	Data management	Legacy systems
Pains	Data management	Legacy cost
Pains	Data management	Business resilience

Pains	Data management	Lack of leadership around open data
Pains	Data management	Security (Business compliance)
Pains	Data management	Timings of when reports are run: 3rd week of the month for data up to previous month end
Pains	Data management	Data protection
Pains	Data management	Fixing reporting errors - is data allowed to change
Pains	Data management	Protecting / masking data
Pains	Data management	Issue: no resource
Pains	Data management	Privacy Shield / Safe harbour (UK / EU)
Pains	Data management	Data overload ('Wheat from Chaff')
Pains	Data management	Documenting the definitions of data
Pains	Data management	Making different datasets consistent and compatible
Pains	Culture	Integrating data into the working day
Pains	Culture	Cultural problems - don't want to move away from XLS or print mentality
Pains	Culture	Expectations of implementing statistical Viz / comparisons which are 'dodgy'
Pains	Culture	F.U.D. (fear, uncertainty and doubt)
Pains	Culture	Cultural (historical)
Pains	Culture	Data used to "beat up" data providers - loss of their support
Pains	Culture	Keeping local copies of other people's data up to date
Pains	Culture	Most people don't care about data or stats

Appendix 3 - Drawnalism output

All workshop output is stored in Google Drive:
https://drive.google.com/open?id=0B8fjAQ6O5B8OUDV0RDdVNnV5aWc

Images shared in advance:
https://drive.google.com/open?id=0B8fjAQ6O5B8OdlptY004RDJhSlE
Suggested pain reliefs



Summary of the tips suggested by Leigh Dodds on the ONS Digital blog