**Discovering the GSS data landscape - Executive Summary**

**OVERVIEW**

An extensive discovery project was undertaken to learn more about the provision of official statistical data across the GSS and to investigate opportunities for development. The following paper is a report of the findings of this project which were submitted to the ONS Design Authority in April 2017.

**GOALS**

1. Produce a comprehensive list of all official statistics/data published and publishable across the GSS.
2. Identify and understand all statistical and data dissemination solutions across the GSS.
3. Work with a small number of departments to understand the capability, limitations and opportunities for developing a transformation pipeline to 4\* open data and the APIs needed to access this data for a presentation layer.

 **KEY ACTIVITIES**

* carried out seven immersive sessions with a range of departments to understand their data journey
* discovery and validation of all statistical outputs by department
* a visualisation of data across departments to show links between datasets
* extensive research across a wide range of users, NSIs, parties working in the data arena, journalists, etc
* development of a proof of concept data pipeline to 4\* open data
* example data visualisations using linked data from the proof of concept
* a large-scale survey of departments to understand current dissemination practices, costs, capabilities and challenges
* a final report and associated recommendations submitted to ONS Design Authority
* socialisation of the findings with interested parties and those who have given their time and resource to the project

**CONCLUSIONS**

* The GSS needs a consistent approach to data dissemination for reuse.
* Internal users need better access to statistical data.
* A solution is needed that can be achieved with existing resources and capabilities in the short term.
* A longer term approach is needed to increase skills and capabilities across the GSS.
* Greater consistency in standards and harmonisation is required to enable more mature formats to be developed as user needs become more sophisticated.
* Users want a less disjointed journey to data, a single access point to data and better formats so they can join data together to give better insight.
* Poor and inconsistent metadata/standards/descriptions are a problem for users.
* There is a significant reliance on 3rd party suppliers to manage dissemination service, adding to the already confusing landscape and a significant cost across departments; circa £24 million per annum.

 **RECOMMENDATIONS**

1. Generate a set of criteria for GSS organisations to produce datasets to a minimum 4\* open data rating, and build an enterprise scale set of transformation pipelines to offer this data to users.
2. Create a machine to transform GSS datasets into a minimum 4\* open data and deliver this via a searchable online data registry. An online facility which adds a level of relevancy and understanding about what the dataset contains. It offers a level of standardization which will help to show the relationships between datasets, their commonalities and their differences.

Our recommendation is undertake number 2 above in the first instance. This will kick-start the move to better data provision showing improvements quickly. There are lots of changes taking place across departments, many undergoing technology refresh programmes which may make it difficult to undertake this change at the same time. Also, we have limited success when trying to make wholesale changes across the system, this is an opportunity to lead by example and iterate as we develop the data system. The key will be in making a big enough bang to act as a catalyst to change rather than a big bang which we then don’t manage well and will only serve to scare those departments who are struggling with this type of work.

Over the next 2 years a gradual shift to number 1 above should be encouraged and supported until all data is produced to appropriate standards.

**WHAT NEXT?**

* 1. If we believe this is the way forward for official statistics then we need a mandate to make it happen. It won’t work without the support of the National Statistician and HoPs across the GSS.
	2. We need to have a high level acknowledgement that to be forward-thinking we need to support data transformation over the current focus of data dissemination.
	3. A steer on priorities for transformation would be helpful. Which themes are seen as key to include in the next phase. (It is worth noting that if this new pipeline was in place the work currently being undertaken by the Racial Disparity Unit could have been simplified).
	4. Funding. The discovery project was funded by ONS, however, there was no commitment to take the next phase forward, and whilst the Design Authority supported the work there is no funding available from ONS.

**Annex 1 - Challenges and opportunities for the GSS**

There are notable challenges to improving the discoverability, access and interoperability of GSS statistical data for users. There are also many opportunities for overcoming these and transforming the way government ‘does data’.

Statistical data is made available through multiple channels and offered in many formats. These contrasting and conflicting services put an unnecessary burden on the user to find and make use of the data.

As a community there is need to examine how to align data production methods to extend the reach and reuse of data. There is a lot to change and no quick fix to do this. We can overcome most of these problems in the short term with the help of technology, but this should be considered only as a catalyst to support data transformation across government.

This annex sets out the problems we have uncovered through our discovery work.

**1. Harmonisation**

Interoperability is an essential element to ensuring we can fulfil the Better Statistics, Better Decisions agenda.

Common labelling of terms used in our datasets are a necessity for user understanding. Many of these terms are the same but are not applied consistently. The term “Gender” is a good example. Some datasets use Sex and some use Gender, but in general they are used indiscriminately. Gender and Sex actually mean different things, but this does not seem a factor in how they are applied.

So you may see:

* Men
* Male
* Boy
* Women
* Female
* Girl

This is a simple example but is common with more complex terms too.

If common terms aren't used across statistical datasets, how do users know which ones share the same meaning? This also applies to internal data sharing between organisations, where terms or IDs are not common across their platforms. Adjusting identifiers for their own use is a burden on an organisation’s resources.

**The challenge:**

* agree a common set of terms across official statistics and use these consistently

**Opportunities:**

* make use of technical standards like RDF to overcome some of these differences by instructing how those terms refer to the same concept
* reduce the reliance on a technological solution through encouraging statistical data producers to conform to the RDF standards, which will reduce maintenance costs in the longer term
* discuss with Methodology the best way to achieve this goal without impacting on statistical integrity
* use agreed common terms to help lower maintenance costs, rather than using an overly complex relational model

**2. Applying Standards**

Many standards exist but there is evidence that these are not used in a consistent way.

**2.1** Classification/code lists: many of these exist, some are official and some are not. Official lists are not always used consistently both within datasets and across the wider landscape; this affects the interoperability of the data.

**2.2** Metadata standards: established standards like DCAT and Dublin Core already exist but aren’t used widely. Applying these standards would aid the discoverability of statistical data and provide users with a common understanding.

**2.3** Structure: the structure of the data is important for reuse too. A generic data schema is not necessarily a good idea as one size doesn't fit all. Following dataset structure definitions like [schema.org](https://schema.org/) could be a starting point to producing consistent datasets, and would improve discoverability within Google, which is the starting point for nearly all search queries around data.

**2.4** Technology standards: building APIs to a common standard may encourage developers to build services on top of our data. There is evidence from a recent API survey carried out by ONS that the APIs we build are difficult to understand and use. Multiply this by the number of organisations providing similar APIs, and this adds up to a large problem for this user group.

 **Challenges:**

* agree and implement technical standards for both APIs and URIs
* introduce metadata schemas to help ensure users understand what each dataset contains
* agree common vocabulary to be used in datasets and enforce harmonisation

**Opportunities:**

* build on the official registers work that GDS are doing to promote the use of agreed lists or classifications to ensure consistency
* align to schema.org standards to make datasets more discoverable using Google
* support “data on the web” principles through the use of data packages which will connect data and metadata for both human and machine consumption
* work toward [CSV on the Web](https://www.w3.org/TR/tabular-data-primer/) standards that ensure data sets are produced to an agreed technical level to aid both use and reuse

**3. Spatial data**Geography as a common link is used to:

* compare over time, changes in statistical data of a single region or between different regions
* display different data sets in a region
* highlight changes
* show spatial relationships between datasets where the regions are at different levels of the same geographic hierarchy

Geography is complex. Different themes operate on specialist geographic areas, for example, environment data uses the area “National Parks”. These “specialist areas” do not fit into or overlap with more well-known area types, including a Local Authority. This can impact cross-cutting analysis.

Differences in area types across the devolved administrations of Scotland, Northern Ireland and Wales alongside England are a complication in the production of GB and/or UK analysis. There has been greater cohesion since Northern Ireland adopted the standard geography coding framework recently, but the specific area types used do not always fit together which makes small area analysis more difficult.

User insight gained during the development of the ONS website proved the provision of small area and hyperlocal data is high on the list of user requirements. Users state that the annual boundary changes also impact on their ability to carry out trend analysis over time.

The presentation of geographical areas is not consistent within products. Not including an area code can impede re-users who want to take data and plug it into their own GIS tool kits.

The inconsistent use of area names is another difficulty for our users, for example: Bristol City, City of Bristol and Bristol, City Of.

**Challenges**:

* identify ways in which specialist areas can be best fitted into more commonly known areas to aid cross-cutting analysis
* create a statistical framework that simplifies how data produced in Scotland and Northern Ireland can be merged with data from England and Wales
* link any changes in geographical boundaries to make it clear for users
* ensure standard geographic coding is used with consistently presented area names
* ensure the geographical hierarchy used by the dataset is clearly presented within the metadata
* identify custodians of geography standards

**Opportunities:**

* support greater interoperability of area-based data and quicker reuse
* support user needs around time series analysis on small area data

**4. Temporal data**

Showing trends over time is another significant user requirement. However, we burden users by using different time periods and presenting these inconsistently to the user. For example:

* quarter on quarter
* rolling quarters - quarter 1, 2, 3, 4
* month ranges - January to March, etc
* calendar year
* academic year
* school year
* financial year

**Challenges:**

* agree standard time labels within datasets
* improve the metadata that defines the classification of time used within the dataset

**The opportunity:**

* support greater interoperability of temporal data and quicker reuse

**5. Technology**

Improvements in technology can’t easily be predicted. Organisations have to work with the tools that are available to them now. There are many middleware options that can help but the crux of the problem is not always the lack of software or technical skills. Not all organisations have the opportunity to introduce the latest software or have the capabilities in house to adopt new tools. Commercial partners are being used widely across the GSS to fill these gaps, and these can come with significant costs. GDS work on a “share before buying” philosophy, however statistical producers are having to find alternate solutions and are bypassing the established spend controls to serve their users.

**Challenges:**

* reduce reliance on costly commercial partnerships
* recognise the need to build tools that can benefit the whole community

**Opportunities:**

* make substantial efficiencies by sharing tools
* reduce the number of individual contracts - employ economies of scale
* where investment is an issue, support a quicker transition to more mature data production processes
* change the way we do the things we do, why we do them and the way we think

**6. Financial**

Re-engineering systems that can operate over networks, and interact with federated systems to automate the entire collection and production process are expensive. This is likely to deter organisations from investing in such activities.

Prescribing the supply of macro level data and metadata in generic schema will incur costs, whether this is commercial partners charging for the service or existing resource efforts to extract the data into the required shape. If existing efforts meet current commitments to publishing national statistics, there is little incentive to change.

There are many projects underway, or being planned, that will transform the way in which organisations operate.

**Challenges:**

* support the production of common open data formats without further increasing the burden on organisations
* reduce reliance on commercial providers to supply and deliver services

 **Opportunities**:

* make substantial efficiencies by sharing tools - encourage a “share before buying” philosophy
* reduce the number of individual providers - employ economies of scale
* capitalise on existing data transformation projects to embed these new ways of working

**7. Capability**

Technology is an important function but the path to more open formats and interoperable statistical data is more an operational change which does not require expert data scientists or data visualisation experts. Users have needs that relate to events in a product’s lifecycle which are not currently taken into account. This makes it more difficult to produce usable data, essentially because more emphasis is put on the publication rather than the curation and preparation of the data.

**Challenges:**

* build a solid community of good practice around producing structured data and metadata by default
* ensure all GSS organisations adopt these as part of their business as usual processes

**Opportunities:**

* accelerate the development of transformation pipelines - manage the process for converting spreadsheets into open data formats
* establish a working forum to raise awareness of new methods/tools, and share problems for advice/resolution
* change the way we do the things we do, why we do them and the way we think
* understand difficult parts of the processes, as a catalyst to solving them

**8. User Needs**

To remain relevant, the GSS needs to be proactive in understanding and adapting to changing user needs. It is impossible to meet or forecast changing requirements in data formats if we don't engage regularly with users of the data. Otherwise, decisions on data provision are based on feedback from old consultations or anecdotal evidence on user needs. This will likely lead to missed opportunities or allocating resources to the wrong areas. This user research needs to be done with as wide a pool of users as possible, and not simply the main stakeholders to ensure validity.

The proof of concept (PoC) will produce a linked data offering (of 5\* data rated output formats). These will enable various output types to be generated including Excel, as a by-product of having well structured data and metadata. The assumption that RDF and Linked Data are the right approach should be challenged through further user research to codify the position. Regardless of this, linked and structured data in a common form will offer huge benefits and will enable us to adapt to user requirements as they change in the future.

**Challenges:**

* understand what users want from statistical producers and products to drive production
* develop the right products to fulfil user needs
* ensure that user research is seen as integral to producing high quality statistical data

**Opportunities:**

* carry out an audit of existing user research across the GSS
* establish a regular user research cycle to ensure user feedback is relevant and current

**9. Cultural**

We shouldn’t continue to rely on data owners and publishers to manually wrangle data into a shape that satisfies dissemination protocols. There must be a change in our data management practices ending this extra burden on teams, by making efficient use of middleware to automate and handle the process of quickly and reliably transferring data to the web.

We also need to refresh our thinking as statistical publishers away from “print mentality”, and instead seeing statistical data as digital products that suit both people and machines.

**Challenges:**

* break out of a “publication model” to let users find data about specific things across datasets, eg for a specific geographical area, industry, etc
* encourage the standardisation of statistics by making it easier to reuse and share codes across organisations, and by highlighting overlaps

**Opportunities:**

* carry out a wider education piece across the GSS with open data experts or developer communities that use our data to discuss future dissemination practices
* outline the benefits of producing machine readable assets that can serve a multitude of audiences, and extend the reach of our data

**10. Future provision of GSS statistical data for reuse**

There is likely to be considerable feeling about how a statistical data service could be provided. Generally the default position is to view this as a website or some kind of web portal. Initially this seems like a perfectly sensible approach as these are commonly understood concepts.

If we are to develop a truly integrated data service, there are many considerations that might mean these are not the approaches to take, such as ownership, governance and the responsibilities of managing such a service, etc. We would need to consider how the service might be in direct competition with other platforms like GOV.UK and data.gov, as well as the implication for the impartiality of organisations such as ONS and the devolved administrations.

There are currently mooted options that may avoid us investing in a monolithic platform and the additional maintenance costs. Data.gov.uk is aiming to provide hosting capabilities for open and linked data in the future. Taking advantage of this service, we could build a suite of transformation pipelines that can pull in newly published datasets in different formats. These could then be delivered as restructured RDF formats to a repository. The repository would have a search and query facility to enable re-users to discover the outputs, query across multiple datasets as they will be provided to the same technical standard, and download data in any specific format. There are costs involved in this as the transformation pipelines would need to be set up and managed in the future but these would be significantly less expensive than a developing a brand new website.

**The challenge:**

* to align the data produced by GSS to enable better use and reuse in terms of findability and format

**Opportunities:**

* adopt common standards developed by schema.org to enable the delivery of datasets on the web ie via a Google search
* enable a significant advancement of the structured open data formats across the GSS
* offer better access to the data and reusability of data across a range of user personas
* reduce the reliance on commercial services to provide data wrangling or dissemination services
* reduce burden on organisations producing handcrafted Excel spreadsheets as these are autogenerated

 **11. Other considerations**

* we are yet to understand the full impact of leaving the EU on statistics and data
* legal and commercial requirements may restrict how we interact with the data
* it has proved challenging in previous projects to agree a single set of standards across the GSS
* more efficient data sharing through the adoption of common practices will reduce effort and time spent on applying codes or labels across organisations
* huge opportunity to work together to reduce production and dissemination effort across the GSS
* there is untapped insight that is currently dormant due to the inability to easily to see what is available across output departments
* *“The best things that will be done with our data, will be done by somebody else” - Open Knowledge.* Our data need to be in better shape so that re-users can add their skills and insight to enlighten debate and influence policy

**Annex 2 – User research**

 **APPROACH**

1. **Deep dive visits to other government organisations**

Organisations were selected because of their varied department sizes, wide range of data subjects, data maturity, and organisational designation. Visits were attended by the project team and covered:

* types of data produced
* data life cycle(s)
* user needs
* existing and upcoming projects around data dissemination
1. **Developed user groups**

Media, academics, charities, pressure groups, general public, OGDs/LAs/councils and commercial users.

1. **Produced and disseminated surveys**
* the main aim of surveys was to provide a list of users willing to be contacted. There was an expectation that the deep dive visits would provide a user base to start with, but organisations had extremely limited user research
* surveys were posted in multiple locations with the aim of specifically targeting identified user groups - media/data journalists ([journalism.co.uk](https://www.journalism.co.uk/)) and existing data users ([GOV.UK](https://www.gov.uk/))
* an additional survey was also sent to known ONS data users as these are known heavy data users
* data.gov.uk also ran a survey on data usage and their user researcher was happy to share their results
* 17 telephone interviews were conducted with users across the full range of user groups
* 2 face-to-face interviews with one enquiring citizen looking for government data for the first time and one with an expert user who works in the open data field
* produced a user journey diagram based on findings (attached)

**FINDINGS**

* all users stated that their journey would be improved if there was a single place to find, or search for all government data
* only one user we spoke to was interested in having commentary that accompanies data, and there was a general need to get to the data quickly
* most users looking for data in a professional capacity have built up experience/ knowledge of where they can find data over time - usually from more experienced colleagues or over the course of their role
* when experienced users with prior knowledge aren’t able to find the data they need, then they rely heavily on Google search
* general public users don’t often have prior experience of sourcing government statistics, so their starting point is generally Google search, (unless they are familiar with the ONS or GOV.UK as sources of statistical information)
* users visit a wide variety of websites and databanks to source their data, often depending on their area of expertise - most commonly visited sites were ONS, devolved administrations and GOV.UK
* the data.gov.uk survey shows that two thirds of users looking for data on their site didn’t find what they needed
* the data.gov.uk survey also showed that 20% of respondents were looking for data not actually provided on their site
* naming of datasets, data categorisation and the taxonomies used across government data websites make searching for datasets difficult for all user groups
* without knowing the name of the data set or the organisation that produces it, users find it more difficult to get to the right data

Data is used in many different ways, depending on the type of user. We have generalised here, rather than stick rigidly to the ONS website personas, as the GSS feel these are not always representative of their users.

Data user goals are:

* generating data intelligence by analysing data
* combining datasets to produce enhanced analysis
* producing a report
* producing their own visualisations (charts, maps, infographics, etc)
* promoting information via social media

It is worth noting that users have explicitly said that they do not require the ability to transform data on a website. They prefer to take the data away and do their own analysis within their own systems.

Also the inconsistencies around metadata and caveats are a barrier to users easily using government statistical data - extra time is needed to find out what data can and can’t be used for, and to do additional cleaning or analysis before datasets can be combined.

**RECOMMENDATIONS**

For the alpha stage of the project, we propose the following user research should be carried out:

* continued research with users from all user groups - these should ideally take the form of face to face observations in the normal working environment
* work closely with FullFact - as a heavy user of government data they are keen to be involved in the project as a critical friend
* further validation of existing knowledge and findings - including the user groups, user needs, pain points
* test proof of concept/prototype with users to get feedback



**Annex 3 - Case study snapshot**

During the pre-discovery phase, we met with 14 organisations to discuss their current data processes to see if they were a good fit for our discovery. In order to get an even spread of organisational structure, topic and data maturity, we developed a set of criteria to measure the suitability of the organisations. The final organisations we worked with were Ofsted, DWP, Scottish Government, Natural England, NHS Digital, Office for Rail and Road, and the Department for International Trade. We spent 2 days visiting each organisation and discussing their data storage, curation, analysis and dissemination processes.

Following our visits, we produced detailed case studies and from these we were able to identify a number of commonalitiesand differences. (The detailed case studies are available and will be shared with the GSS during the socialisation of the report).

**CURATION AND WRANGLING**

**Commonalities**

* All organisations used consistent toolkits across each organisation, such as SAS and Excel
* Most organisations felt metadata was not a priority and often applied it after the data had been prepared, and often applied it inconsistently
* Most organisations have security and/or legal constraints around producing their data
* Most organisations can extract structured data from their databases
* Most organisations have data transformation projects forthcoming or currently happening, but were also clear that they had the skills and capabilities in house to produce their current stats offerings
* Most organisations found getting access to other organisation’s data problematic
* Most organisations are working with a wide variety of data suppliers, including private and commercial organisations
* Some organisations would be impacted directly by the upcoming departure of the UK from the EU

**Differences**

* Most organisations use disparate geographies that don’t interoperate with the geographies of other organisations
* Most organisations had different approaches to storing data
* Most organisations use different standard code lists which makes harmonisation difficult

**PUBLISHING**

**Commonalities**

* All organisations had at least one bespoke publishing solution, which was developed and/or supported by a third party vendor. The results of our survey show that this sample was representative of the rest of the GSS.
* All organisations had the requirement to publish on the GOV.UK release calendar, and 5 of the 7 were organisations who are native to GOV.UK.
* All organisations use a range of formats across their own published data.
* All organisations collected some data that was only used internally, and not released to the public.
* All organisations had very limited user research and user engagement work. Any information they did have was based on either the needs of internal users, or from consultations and surveys done years before.
* All organisations had a limited use of metrics and analytics to inform or drive ongoing work around statistics.
* All organisations felt that data visualization was a user requirement.
* Most organisations had known user groups and well-known pressure groups that relied on their data.
* Most organisations had currently unpublished data that they felt could be made public.
* Some organisations published to data.gov.uk, but none solely to the site.
* Some organisations were unclear exactly where they published.
* Some organisations had users and resellers which they were aware of, but some were unaware of who may be repackaging their data and republishing it.
* Very few organisations had a legal gateway for researchers to access unpublished data.

**Differences**

* All organisation’s standards and release formats are different to each others

**Annex 4 - Project due diligence**

We set up this discovery project not to solve a single problem, but to better understand the current statistical landscape and to identify the drivers and opportunity for change across our services. There is a wealth of information and resources in the statistical community that we could call upon. The knowledge we gained enabled us to set the right direction for the project, and it provided a solid platform to begin from. The information contained in this annex shows the research we have undertaken, and should provide a level of confidence in the findings of this report.

**1. Methods**

In the course of this project, we have:

* undertaken 2-day deep dives with candidate departments: DWP, NHS Digital, Scottish Government, Office for Rail and Road, Department for International Trade, Ofsted and Natural England. These were detailed exploratory sessions working with the teams who handled statistical production and dissemination activities in each organisation.
* produced a detailed questionnaire and sent it to organisations via the HoPs to confirm dissemination procedures across the GSS, to supplement the deep dive information
* confirmed the provision and periodicity of statistical releases with organisations to gain a definitive annual list of outputs to build our “solar system of statistics”
* produced user research surveys to elicit basic user feedback and to recruit users willing to take part in interviews - to get coverage of all user groups, versions of the survey were available via GOV.UK, Journalism.co.uk, TheConversation.com/uk and sent to known ONS data users groups
* undertaken telephone interviews with people across each of the identified user groups
* Observed, (through face-to-face interview), two people trying to find government data

**2. Research**

We were also careful to understand what work was already in progress, or had been carried out, within government and the wider statistical community, both in the UK and internationally. Through this, we were able to contact potential collaborators, identify any similar approaches, and learn lessons from other projects.

We used a number of papers and blogs as background information to draw upon expert knowledge, from:

* [GDS](https://gds.blog.gov.uk/)
* [European OpenCube project](http://opencube-toolkit.eu/case-studies-2/)
* [INSEE](https://www.insee.fr/fr/information/2410988)
* [Data.Gouv.fr](https://www.data.gouv.fr/)
* [Eurostat work on Linked Data](http://eurostat.linked-statistics.org/) and [metadata](http://ec.europa.eu/eurostat/data/metadata/metadata-structure)
* [Open Knowledge Foundation](https://okfn.org/projects/)
* [Data.World](https://data.world/)
* [Defra Digital](https://defradigital.blog.gov.uk/)
* [Open Government Intelligence project](http://www.opengovintelligence.eu/)
* [Open data in Health Sector report](http://openhealthcare.org.uk/open-data-in-the-health-sector/)

We had discussions with the following organisations:

* National Audit Office
* Open Data Institute
* NISRA
* British Medical Journal
* Google
* Welsh Government
* many expert resources within ONS

These conversations provided useful insights into the needs of users, and what was happening across government and commercial sector. This helped to shape our decisions around the methods we wanted to use.

We also attended a number of sessions and meetings. We listened to people who had previously produced data portals, or who wanted to gain a wider insight into producing data portals from federated systems and setting standards in the open data space.

These were:

* the International [Open Data Conference](http://opendatacon.org/) in Madrid
* sessions with [data.gov.uk](https://data.blog.gov.uk/category/data-infrastructure/data-gov-uk/) - they are undertaking a similar project to ours but focusing on data providers and management/transparency information. We’ve worked together to share and learn from each others experiences
* pre-discovery conversations with DCMS, DCLG, DfE, HMRC and PHE - capturing information to formulate what criteria we would use to select deep dive departments.

**3. Socialisation of the Discovery**

We’ve worked hard to socialise the project by attending:

* GDS Product Team/GOV.UK workshops - presenting the approach we were taking and seeking further advice and validation
* Open Data Conferences - [Open Data Camp](http://odcamp.org.uk/) was a big open data community event which offered up opportunities to learn and talk about the work we are doing
* several GSS Presentation and Dissemination Committee and GOV.UK sub-group meetings
* Race Disparity Unit workshops - to help their team understand the work and the challenges faced on a project like ours, which has significant similarities to their work
* international conferences - the approach was presented at a meeting in Eurostat to over 50 NSIs and commercial partners
* external user group sessions including Publish My Data and the cross-government Publishing Community Group - presenting the work we had done to a wider audience of data publishers to seek challenge from an expert group
* internal ONS user groups - the National Statistics Harmonisation Group and the Data Policy and Standards Group, both major stakeholders in this work
* ESS Linked Open Data [Workshop in Malta](https://ec.europa.eu/eurostat/cros/content/ess-workshop-dissemination-official-statistics-open-data_en) included many open data advocates from across Europe indicating how much of a focus this is across NSIs

We have also published 7 blog posts on the GSS website to keep people informed of our progress:

* Blog 1: [You don’t know what you don’t know - the GSS Data Discovery](https://gss.civilservice.gov.uk/blog/2016/09/dont-know-dont-know-gss-data-discovery/)
* Blog 2: [Getting stuck in - what we’ve learnt so far](https://gss.civilservice.gov.uk/blog/2016/12/getting-stuck-weve-learnt-far/)
* Blog 3: [Our 3 Kings: a prototype, a questionnaire but where’s the data architect?](https://gss.civilservice.gov.uk/blog/2016/12/3-kings-prototype-questionnaire-wheres-data-architect/)
* Blog 4: [Our tour of the North](https://gss.civilservice.gov.uk/blog/2017/01/discovering-gss-data-landscape-tour-north/)
* Blog 5: [Where we are so far](https://gss.civilservice.gov.uk/blog/2017/01/discovering-gss-data-landscape-tour-north/)
* Blog 6: [Trade, trains and automobiles](https://gss.civilservice.gov.uk/blog/2017/02/trade-trains-automobiles/)
* Blog 7: [Contemplate, germinate, formulate](https://gss.civilservice.gov.uk/blog/2017/03/contemplate-germinate-formulate-gss-discovery-project-comes-next/)

Further to this, we have developed a communication strategy for the promotion of the final report, and to show the proof of concept to a wider audience.

**Annex 5 - Project roadmap**



# Annex 6 - Discovering the Statistical Landscape of the GSS survey

| **Where is statistical data stored in your organisation?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Data warehouse |

|  |
| --- |
|   |

 | 54.55% | 24 |
| 2 | Private cloud |

|  |
| --- |
|   |

 | 34.09% | 15 |
| 3 | Internal database |

|  |
| --- |
|   |

 | 72.73% | 32 |
| 4 | Desktop |

|  |
| --- |
|   |

 | 15.91% | 7 |
| 5 | Other (please specify): |

|  |
| --- |
|   |

 | 25.00% | 11 |

| **What analytical tools do you use to validate and analyse your statistical data?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | SAS |

|  |
| --- |
|   |

 | 40.91% | 18 |
| 2 | SPSS |

|  |
| --- |
|   |

 | 65.91% | 29 |
| 3 | Excel |

|  |
| --- |
|   |

 | 93.18% | 41 |
| 4 | R |

|  |
| --- |
|   |

 | 56.82% | 25 |
| 5 | Other (please specify): |

|  |
| --- |
|   |

 | 52.27% | 23 |

| **In what format(s) do you prepare and manage your statistical data for external dissemination?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Excel |

|  |
| --- |
|   |

 | 100.00% | 44 |
| 2 | PDF |

|  |
| --- |
|   |

 | 86.36% | 38 |
| 3 | Word |

|  |
| --- |
|   |

 | 56.82% | 25 |
| 4 | ODS |

|  |
| --- |
|   |

 | 25.00% | 11 |
| 5 | JSON |

|  |
| --- |
|   |

 | 9.09% | 4 |
| 6 | RDF |

|  |
| --- |
|   |

 | 9.09% | 4 |
| 7 | CSV |

|  |
| --- |
|   |

 | 61.36% | 27 |
| 8 | XML |

|  |
| --- |
|   |

 | 13.64% | 6 |
| 9 | Other (please specify): |

|  |
| --- |
|   |

 | 40.91% | 18 |

| **How do you publish your statistical data online?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | CKAN |

|  |
| --- |
|   |

 | 13.64% | 6 |
| 2 | Whitehall Publisher |

|  |
| --- |
|   |

 | 45.45% | 20 |
| 3 | DKAN |   | 0.00% | 0 |
| 4 | Socrata |   | 0.00% | 0 |
| 5 | PMD |

|  |
| --- |
|   |

 | 4.55% | 2 |
| 6 | Tableau |

|  |
| --- |
|   |

 | 13.64% | 6 |
| 7 | Other (please specify): |

|  |
| --- |
|   |

 | 59.09% | 26 |

| **Which open data rating does your organisation comply with?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | 1\* PDF |   | 0.00% | 0 |
| 2 | 2\* Excel |

|  |
| --- |
|   |

 | 27.27% | 12 |
| 3 | 3\* CSV (or other) |

|  |
| --- |
|   |

 | 59.09% | 26 |
| 4 | 4\* RDF |

|  |
| --- |
|   |

 | 2.27% | 1 |
| 5 | 5\* Linked |

|  |
| --- |
|   |

 | 4.55% | 2 |
| 6 | No rating |

|  |
| --- |
|   |

 | 6.82% | 3 |

| **Do you provide open public APIs for users to access your statistical data?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 25.00% | 11 |
| 2 | No |

|  |
| --- |
|   |

 | 75.00% | 33 |

| **Do you provide technical standards/guidance to inform users how to use these?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 81.82% | 9 |
| 2 | No |

|  |
| --- |
|   |

 | 18.18% | 2 |

| **Is your organisation planning or currently undergoing IT projects that will change how you store, curate and disseminate statistical outputs in the next 12 months?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 75.00% | 33 |
| 2 | No |

|  |
| --- |
|   |

 | 25.00% | 11 |

| **We are trying to understand the value of third party vendors who provide dissemination services, such as: data hosting - external cloud or data centres data cleansing and publishing activities website hosting and maintenance costs. Can you advise on the total annual costs associated with these activities?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | <£100k |

|  |
| --- |
|   |

 | 52.27% | 23 |
| 2 | <£250k |

|  |
| --- |
|   |

 | 18.18% | 8 |
| 3 | <£500k |

|  |
| --- |
|   |

 | 6.82% | 3 |
| 4 | <£1m |

|  |
| --- |
|   |

 | 11.36% | 5 |
| 5 | >£1m |

|  |
| --- |
|   |

 | 4.55% | 2 |
| 6 | >£3m |

|  |
| --- |
|   |

 | 2.27% | 1 |
| 7 | >£5m |

|  |
| --- |
|   |

 | 4.55% | 2 |

| **Are you currently working to develop technical expertise around data in your organisation?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 90.91% | 40 |
| 2 | No |

|  |
| --- |
|   |

 | 9.09% | 4 |

| **If not at this time, what are the drivers for this?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Abundant skills in house |

|  |
| --- |
|   |

 | 25.00% | 1 |
| 2 | Not an organisational priority |

|  |
| --- |
|   |

 | 50.00% | 2 |
| 3 | Financial constraints |

|  |
| --- |
|   |

 | 25.00% | 1 |
| 4 | Use of third party vendor |   | 0.00% | 0 |
| 5 | Other (please specify): |

|  |
| --- |
|   |

 | 25.00% | 1 |

| **Which of the following areas are you focussing on?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Data visualisation |

|  |
| --- |
|   |

 | 82.50% | 33 |
| 2 | Open data |

|  |
| --- |
|   |

 | 65.00% | 26 |
| 3 | Data journalism |

|  |
| --- |
|   |

 | 20.00% | 8 |
| 4 | Data science |

|  |
| --- |
|   |

 | 82.50% | 33 |
| 5 | Other (please specify): |

|  |
| --- |
|   |

 | 22.50% | 9 |

| **How are you developing these skills?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Recruitment |

|  |
| --- |
|   |

 | 56.10% | 23 |
| 2 | Upskilling |

|  |
| --- |
|   |

 | 100.00% | 41 |
| 3 | Knowledge sharing |

|  |
| --- |
|   |

 | 80.49% | 33 |
| 4 | Other (please specify): |

|  |
| --- |
|   |

 | 9.76% | 4 |

| **Does your organisation have enough of the skills and resources needed to drive forward your open data agenda?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 43.18% | 19 |
| 2 | No |

|  |
| --- |
|   |

 | 56.82% | 25 |

| **Have you done any user research or gathered feedback to help identify user needs/requirements around the provision of statistical data?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 77.27% | 34 |
| 2 | No |

|  |
| --- |
|   |

 | 22.73% | 10 |

| **Have your users commented on the usability of the GOV.UK platform for statistical data?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 31.82% | 14 |
| 2 | No |

|  |
| --- |
|   |

 | 68.18% | 30 |

| **Do you use a spreadsheet package to carry out any data modelling?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 65.91% | 29 |
| 2 | No |

|  |
| --- |
|   |

 | 34.09% | 15 |

| **Do your teams use standard code lists/registers/classifications when publishing your statistical data?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 72.73% | 32 |
| 2 | No |

|  |
| --- |
|   |

 | 27.27% | 12 |

| **Have your teams experienced any barriers to providing data in open formats?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 63.64% | 28 |
| 2 | No |

|  |
| --- |
|   |

 | 36.36% | 16 |

| **Which of the following barriers have they experienced?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Technical |

|  |
| --- |
|   |

 | 75.00% | 21 |
| 2 | Financial |

|  |
| --- |
|   |

 | 32.14% | 9 |
| 3 | Skills capability |

|  |
| --- |
|   |

 | 53.57% | 15 |
| 4 | Support |

|  |
| --- |
|   |

 | 28.57% | 8 |
| 5 | Other (please specify): |

|  |
| --- |
|   |

 | 57.14% | 16 |

| **Do you apply metadata standards to your statistical data releases?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | StatDCAT |   | 0.00% | 0 |
| 2 | DCAT |

|  |
| --- |
|   |

 | 6.82% | 3 |
| 3 | Dublin Core |

|  |
| --- |
|   |

 | 11.36% | 5 |
| 4 | Other tagging |

|  |
| --- |
|   |

 | 13.64% | 6 |
| 5 | No |

|  |
| --- |
|   |

 | 75.00% | 33 |

| **Do you publish your data on GOV.UK?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | All |

|  |
| --- |
|   |

 | 36.36% | 16 |
| 2 | Some |

|  |
| --- |
|   |

 | 38.64% | 17 |
| 3 | No |

|  |
| --- |
|   |

 | 25.00% | 11 |

| **Do you publish your data on data.gov.uk?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | All |

|  |
| --- |
|   |

 | 11.36% | 5 |
| 2 | Some |

|  |
| --- |
|   |

 | 43.18% | 19 |
| 3 | No |

|  |
| --- |
|   |

 | 45.45% | 20 |

| **Do you publish data on your own website or other platform?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 68.18% | 30 |
| 2 | No |

|  |
| --- |
|   |

 | 31.82% | 14 |

| **Do you have publishable data that is not currently published?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 75.00% | 33 |
| 2 | No |

|  |
| --- |
|   |

 | 25.00% | 11 |

| **Do you offer access to data using a special licence or legal gateway?**  |
| --- |
|  | **Response Percent** | **Response Total** |
| 1 | Yes |

|  |
| --- |
|   |

 | 47.73% | 21 |
| 2 | No |

|  |
| --- |
|   |

 | 52.27% | 23 |

**Annex 7 - Simple infrastructure workflow**

This annex illustrates the high level approach taken to transforming differing output formats into RDF and Linked Open Data, in the discovery phase. (This is provided as information only and should not be considered a solution).

The reason for developing a prototype was to illustrate what is potentially a difficult concept to fully grasp in a written report. It also provides a tangible view of how we can take advantage of this technical standard to offer greater utility of data to users.

In essence there is a very basic problem - data is produced in different formats within an organisation, and is also differently to other organisations. We need a solution able to manage this variability.

1. We chose to use “DataBaker”, a tool familiar to our team, which uses [Python](https://en.wikipedia.org/wiki/Python_%28programming_language%29) scripts to deconstruct tabular formats into single generic machine-readable content.
2. Data from more sophisticated systems can bypass the deconstruction stage, as the data may be called on through machine to machine services.
3. “Grafter”, a tool provided by the project’s commercial partners Swirrl, takes the deconstructed data and transforms it into RDF. It also builds the triples that describe the relationships between the data.
4. The resulting datasets are published to Swirrl’s [PublishMyData](http://www.swirrl.com/) Platform for validation by the team to ensure the data has been transformed correctly. The platform is the basis for users to [search, query and interact with the data](https://public.tableau.com/profile/jamie.whyte#!/vizhome/ONS-Linked-Data-Pilot/Story2).

Each organisation could have its own instance of PublishMyData, or could share instances where appropriate. The high level architecture is flexible and offers the possibility to distribute data management and storage.

The API and query facilities allow the data and metadata in the system to be used flexibly by other systems, including websites, search tools, applications or visualisations. Standards-based APIs and standards-based ways of representing the data make it possible to combine data from more than one publisher and/or more than one instance of PublishMyData. This approach therefore enables a consistent, authoritative, perhaps distributed, collection of data to support a range of dissemination approaches.

It is worth noting that in the recent Open Data Watch, [Open Data Inventory 2016](http://odin.opendatawatch.com/report/rankings) the UK ranked 46 in the world. This could be attributed to two things, firstly that we have limited open data across government, and secondly that our disparate solution to disseminating data makes it difficult to properly measure.



**Annex 8 - Solar System of Statistics**

One of the primary purposes of the discovery is to test whether a single mechanism for discovering GSS data is an actual user need.

Alongside the request to complete the questionnaire (see Annex 5) organisations were asked to review a list of the statistical outputs currently listed on the GOV.UK release calendar. The project was keen to understand whether this list is accurate reflection of what is produced. The intention was to have a definitive list of national and official statistics that will be the core of a “GSS Data Registry”.

The list will be used to populate the proof of concept and will be fully indexed to enable a powerful search capability across the available data sets. We’ll ensure these are evaluated through user testing.

A secondary purpose was to gain a holistic view of the statistical landscape. The image added on the next page is a snapshot of an interactive linked data tool we have developed that provides a connected representation of the GSS data products and the topics they contain. (This version is not a fully populated image but an extract of that view). The function of this tool is to visualise the full extent of datasets available, what information they contain and where connections exist.

Viewing the data in this way allows us to:

* gain a sense of the range of difference terms used when labelling similar things in datasets. This should offer quick wins in identifying where harmonised terms can easily be agreed on across these datasets.
* identify overlaps where different releases are using the same data, so we can reduce user confusion and better understand where gaps exist.

Adding more datasets to the registry will contribute a further level of scrutiny, and hopefully gain further perspectives of the benefits that bringing GSS statistical data together can have.

