Open Data Consultancy

Final Report

Summary

This report presents the outcomes of the Open Data Consultancy study that Swirrl IT Limited was commissioned to carry out for the Scottish Government from September to November 2013.

The Scottish Government
Riaghaltas na h-Alba

‘Open Data’ is data that is accessible to anyone (usually via the internet), in a machine readable form, free of restriction on use. Adoption of this approach to information sharing is growing in the public sector, with anticipated benefits for transparency, efficiency of government and economic growth.

The study included a pilot of the Linked Open Data approach to data publishing, incorporating the Scottish Index of Multiple Deprivation, selected education data and supporting geographical information. Linked Data is a specific web-based technology for data access and data integration, sometimes known as ‘5-star’ data. Linked Open Data is the application of linked data in an open context.

The pilot demonstrated that linked data publishing by the Scottish Government is feasible at modest cost. It provides flexible and powerful mechanisms for producing dynamic visualisations and downloads of data, as well as machine-readable access. The pilot demonstrated how data from different sources can be usefully interconnected.

A process of gathering user feedback has now begun to assess the reaction of data users to the techniques demonstrated in the pilot.

The study analysed what steps would be needed to extend the pilot to the whole of Scottish Neighbourhood Statistics; and concluded that this is feasible.

Linked open data helps make data more discoverable and makes it easier to combine data from different sources. These are challenges that are also encountered in the non-public information environment within the Scottish Government. The study assessed whether the technologies used for open data could also be a good solution for improved data management within the Analytical Services of the Scottish Government. In this context it is important to have control of security and versioning and to protect the investment in existing tooling and skills. Any new system must allow integration with various data analysis software packages.

Linked open data seems promising for internal data management, but is less proven in this context than for open data publishing. Therefore a small scale experiment would provide useful insight to assist a future decision on the suitability of this approach.

Finally, the report assesses what would be required to allow the Scottish Government to use the linked open data approach across all its data publishing activities. It presents a high level view of what such a change in practices would involve and what benefits it might bring.

The objective is more effective exchange of information, both internally and externally: an ambitious goal that will take time to achieve. It will depend on interoperability of many different data-holding systems. To enable that interoperability requires establishing a series of standards: existing open standards from the World Wide Web Consortium combined with the government's own standards and conventions, and supported by reference data sets. Finding the right balance of standardisation and flexibility for innovation should enable a successful ecosystem of data sharing.

This will need to be backed up with new knowledge and skills amongst government staff and a culture and goals where the success of a data curator is measured by how others use their data. The value of open data is created when someone puts it to good use: choices of technology and design of processes should be made with that in mind.

Glossary and abbreviations

5-star open data	Open data that achieves the highest rating on the 5 star scale proposed by Sir Tim Berners Lee for open government data.
API	Application Programming Interface. A set of specifications implemented by software, which allows access to data by external computer programs.
Concept scheme	A maintained list of concept descriptions, typically organised into a hierarchy. Often used to define the possible values of a dimension of statistical data.
CSV	Comma Separated Values. A commonly used file format for tabular data.
Data Cube	A data structure often encountered in statistical data, characterised by a collection of observations sharing a common set of dimensions, measures and attributes. A data table is a two-dimensional data cube.
GIS	Geographical Information Systems: software packages used for management and analysis of geographical data.
HTTP	Hypertext Transfer Protocol. The foundational standard for data communication on the web.
JSON	JavaScript Object Notation. A widely used text-based data format, defined by an open standard. Originally developed as part of the JavaScript language, but now popular for web-based software development in a variety of programming languages.
Linked data	A method for publishing structured data, building on standard web technologies and designed to promote interlinking and interconnecting different data sources.
Linked open data	The linked data method applied to publishing of **open data**.
Ontology	A systematic description of an **RDF** data model, comprising a vocabulary of properties and classes, their definitions and how they are inter-related.
Open data	Data that is accessible (usually via the internet), in a machine

	readable form, free of restriction on use.
Open data approach	The set of processes and technologies associated with the creation and use of open data.
RDF	Resource Description Framework. A data model and associated data syntaxes for representing data as a series of statements about resources. It is the standard format used in **linked data**. It is defined by a series of **W3C** specifications.
REST	Representational State Transfer: an architectural style for software designed to work in a distributed web-based environment.
SAS	A company that produces software for analysis of statistical data. Also used to refer to their statistical analysis product that is widely used in Scottish Government Analytical Services divisions.
SIMD	Scottish Index of Multiple Deprivation
SPARQL	SPARQL is a query language for RDF databases. It is defined by a **W3C** specification.
SPSS	A software package for statistical analysis, used in the Scottish Government Analytical Services divisions.
SQL Server	A relational database system produced by Microsoft and widely used in the Scottish Government.
Triple store	A database designed for storing data in RDF format. RDF represents data by a series of 'triples', each consisting of a subject, a property and an object.
URI	Uniform Resource Identifier. A string of characters following a standard syntax used to give a name to a web resource or real world entity.
W3C	The World Wide Web Consortium: an international community that sets the architecture and standards used on the web.
XML	Extensible Markup Language. A widely used open standard data format.

1. Introduction

This report presents the outcomes of the Open Data Consultancy study that Swirrl IT Limited was commissioned to carry out for the Scottish Government from September to November 2013. The study incorporated 3 main components:

- Understanding the needs of the Scottish Government regarding data publishing and internal data management and advising how the open data approach could be applied in this context
- Holding a series of three workshops to raise awareness and explain the details of open data to staff from the Scottish Government
- Creating a pilot website to illustrate how open data, and in particular linked open data, could be applied in practice to improve access to Scottish Government statistical data.

This report describes the findings and suggestions arising from the study and aims to provide useful reference information and discussion points, to assist the Scottish Government in developing a plan for how it makes use of open data in future.

2. Open data overview

2.1 What is open data?

Open data in a government context can be defined as data that is:

- accessible (ideally via the internet) at no more than the cost of reproduction, without limitations based on user identity or intent;
- in a digital, machine readable format for interoperation with other data; and
- free of restriction on use or redistribution in its licensing conditions[1].

2.2 Why is open data useful?

The prime objective of open data is to get the right information to the people who need or want it, in a form that allows them to use it. The promise of open data is to achieve this quickly and cheaply.

Systematic use of open data is relatively new and the availability of quantitative evidence on cost and impact is still limited. However, a number of early success stories are discussed below.

The audience for government open data is broad: it includes the government itself, other public sector organisations (such as local authorities, public bodies), businesses who use public sector data in their operations, businesses that add value to public sector data and resell it, academic researchers, charities and other civil society organisations, as well as individual members of the public.

The uses for such data are diverse and will vary across these groups. The potential benefits they derive from using the data are often grouped into three broad categories:

1. transparency and accountability of government
2. efficiency and effectiveness of the public sector
3. supporting economic growth

We will consider each of these in turn.

[1] https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential

1. Transparency and accountability

The government is elected by the people and the activities of government are funded by taxpayers: it is therefore a reasonable expectation that the citizens of a country are able to see what their government is doing and how their money is being spent.

The Open Government Partnership, of which the UK is a member, is founded on the basis that more transparent government contributes significantly to the goal of "improving the quality of governance, as well as the quality of services that citizens receive"[2].

2. Efficiency and effectiveness of the public sector

Information sharing via open data is not an end in itself, but rather something that can assist the essential functions of public sector organisations. Making relevant data more accessible and usable has the potential to assist the government to:

- know whether policies are working
- design new policies or adapt and improve existing ones
- target investment
- improve targeting and delivery of services
- enable collaboration between organisations in service delivery
- improve data quality by increasing data visibility and enabling third parties to contribute improvements.

"Scotland's Digital Future: Delivery of Public Services"[3] describes a vision for Scotland where

> *"digital technology provides a foundation for innovative, integrated public services that cross organisational boundaries and deliver to those in most need, and for services for business that promote growth"*

The foreword to this report also illustrates the central role of data in digital delivery of public services. It is highly significant that this vision is one where services *cross organisational boundaries*. Not only do different parts of central government need to share information effectively, but information must also be shared with local government, health services and other public sector organisations – and often also with charities, civil society groups, businesses and individuals.

Recognising the central importance of data, as opposed to the applications or systems that hold it, is a necessary step in escaping from the data silo problems all

[2] http://www.opengovpartnership.org/about/mission-and-goals
[3] http://www.scotland.gov.uk/Publications/2012/09/6272

large organisations face. Making data interoperable and re-usable offers the possibility of large cost savings by reducing search costs, reducing duplication, reducing data processing time and reducing mistakes.

3. Supporting economic growth

Alongside transparency and efficiency of government, the other central objective of the open government data movement is to encourage economic growth. The 2013 Shakespeare Review of Public Sector Information[4] commissioned a Market Assessment[5] carried out by Deloitte. The study analyses the ways in which public sector information is used by different market segments and estimates the economic value arising from use of the information. The study concludes that UK-wide the direct benefits are around £1.8 billion per year and the indirect benefits are around £5 billion per year. A number of barriers to effective use of data are identified – if these are overcome, the overall benefits could be significantly higher.

While the details of the study are informative, the overall implication is a simple one: enabling greater use of government data, both within the public sector and by the private sector will bring important economic benefits. The message for a public sector data owner boils down to: how can I help more people use my data?

2.3 Examples of open data publishing and exploitation

There is a growing portfolio of open data available in Scotland and across the UK. The Scottish Government and other public sector organisations in Scotland publish a large quantity of open data, notably the Scottish Neighbourhood Statistics, but also significant other resources on, for example, education[6], health[7] and the environment[8].

The UK government's main open data site http://data.gov.uk is a central resource for public sector open data in the UK, listing thousands of openly available datasets. This includes data about Scotland, particularly spatial data related to the INSPIRE directive[9] and on non-devolved issues.

There is also a growing collection of local government initiatives around open data, for example:

- The Glasgow Future City project is investigating a number of innovative uses of open data in the context of city management[10]

[4] https://www.gov.uk/government/publications/shakespeare-review-of-public-sector-information
[5] https://www.gov.uk/government/publications/public-sector-information-market-assessment
[6] http://www.scotland.gov.uk/Topics/Statistics/Browse/School-Education
[7] http://www.isdscotland.org/
[8] http://www.environment.scotland.gov.uk/
[9] http://data.gov.uk/location/inspire
[10] http://data.glasgow.gov.uk/ and http://open.glasgow.gov.uk/

- Councils including open datasets on their websites, for example Aberdeen City[11]
- Dedicated city data websites, eg London[12] and Manchester[13]
- Local authority information system or 'observatory' sites, such as KnowFife[14] and the Hampshire Hub[15]

Many of these open data publishing initiatives are based around a catalogue of datasets, providing a list of available data, with descriptive information and a link to where the data can be downloaded. Other initiatives have taken a richer approach using linked data and other types of Application Programming Interface (API) to enable data to be exploited by other applications. Examples include:

- The Office for National Statistics 'NOMIS' API[16]
- The Department for Communities and Local Government 'Open Data Communities' site[17]
- Environment Agency data on bathing waters[18]
- Transport for London live data feeds[19]

It is easier to gather data on the supply of open data than it is to assess systematically how it has been effectively used. Usage data tends to be anecdotal and based on isolated representative examples. Nonetheless, it can easily be seen how these examples could be replicated more broadly. Many examples are of small impact in themselves, but uses of data can be so broad and so diverse that the total impact is significant.

Comparison of spending data between local authorities has helped some to identify where better deals can be had, sometimes leading to shared procurement approaches.

Release of open data on NHS General Practice prescriptions in England enabled Prescribing Analytics[20] to compare the spending on branded versus generic statins across England, concluding that if best practices were applied in all GP practices, around £200 million per year could be saved on the costs of these drugs.

[11] http://www.aberdeencity.gov.uk/open_data/open_data_home.asp
[12] http://london.gov.uk
[13] http://datagm.org.uk
[14] http://knowfife.fife.gov.uk/IAS/
[15] http://protohub.net/
[16] http://www.nomisweb.co.uk/
[17] http://opendatacommunities.org
[18] http://environment.data.gov.uk/
[19] http://www.tfl.gov.uk/businessandpartners/syndication/
[20] http://prescribinganalytics.com/

This analysis could have been carried out by the relevant authorities without use of open data, but it did not take place until the data was available openly: or at least if it did, it wasn't acted upon. Many aspects of open data involve doing existing things better or more cheaply rather than doing things that were impossible before. Most organisations suffer from pressure on time and resources. Lowering the barrier to information access can mean that opportunities that were previously ignored can now be exploited.

In addition to use of open data for large scale cost analysis, there are numerous examples of small scale efficiency improvements and new opportunities created by the existence of high quality open data. The impact in each individual case is small, but if replicated many times, the overall effect can be significant.

Transport for London (TfL) offers a number of APIs to provide live access to bus and train locations and arrival times. This has led to the creation of numerous travel information apps, assisting users in journey planning and avoiding known disruptions. TfL has estimated the impact of the accumulated small savings in journey and waiting times that better travel information has created. They have concluded that this leads to a return on investment of around 58:1 for their open data initiatives[21].

Another important example is the widespread use of online mapping tools, such as Google Maps (which in UK makes heavy use of open data from the Ordnance Survey). The existence of instantly accessible high quality mapping and photography has revolutionised journey planning for businesses and individuals. The time savings may amount to only a few minutes per journey, but applied hundreds of millions of times, this amounts to a notable increase in productivity.

An important consideration, illustrated by some of these examples, is the role of intermediaries in serving the needs of small niche markets. The owner or publisher of open data cannot predict or directly serve the needs of all users of the data. But by making the data openly available in a re-usable form, other business or organisations that understand the needs of a particular market or community can process or combine selected data in order to meet those needs. It underlines the importance in public sector information publishing of allowing commercial re-use of data. By allowing businesses to add value to the data for their customers, it amplifies the overall benefits of making the data available.

[21] http://www.slideshare.net/dirdigeng/20130719-muirfield01

2.4 Costs and risks

While open data has the potential to bring many benefits, there are of course also costs and risks associated with it.

Sharing data in a way that encourages and enables others to use it requires effort on the part of the data publisher: to organise and annotate data and present it in a highly usable form. The process of publishing 5-star open data is explained in detail in Chapter 5 in the context of the pilot. The main additional costs associated with open data publishing arise from the time of well qualified staff in organizing, transforming and presenting their data. Depending on available in-house skills, it may be necessary to bring in external experts to assist.

Publishing open data will lead to a certain amount of cost in software licences and web hosting, but these costs are likely to be relatively modest in comparison with cost of staff time.

Implementing open data publishing at a significant scale will also require a training and skills development programme, which will also have a cost.

Most of the costs in this process lie with the publisher and many of the benefits come to the end user of the data. Since the essence of open data is not to charge for it, this imbalance can sometimes be seen as an obstacle. In a public sector context however, it is important to note that the biggest users of public sector data are often other parts of the public sector. Also, government has a role in promoting the success of the private sector economy, so investment in assisting the public sector as a whole and the economy as a whole is a legitimate and justified approach. Hence costs incurred by the group publishing the data can be offset against overall reduction in the cost of discovery, distribution and exploitation of data across the public sector.

There can be risks associated with open data publication, arising primarily from making data available to people who did not previously have access to it, and having limited influence on what they do with it. Commonly raised risks include:

- there could be mistakes in preparing the data
- users of the data misunderstand what it means or how it can be used
- data quality and defining what purposes the data is valid
- lack of control on how users will use the data
- managing user expectations
- creating more work for data producers, especially if data needs to be published to a strict schedule
- resulting analyses may have unintended political or policy consequences

- protecting interests and privacy of individuals

These are legitimate concerns and the approach to data publication should be designed in order to mitigate these risks. An important concept is the idea of enabling and promoting 'responsible re-use' of data. In essence this involves ensuring that the provenance, meaning and possible limitations of data are clearly documented and easily accessible to potential users of the data. In most cases, mis-representation of data by users is not malicious and the risk of it can be reduced by clear communication. In cases of deliberate mis-representation of data, the data owner can point interested parties to explanatory material and ensure that sufficient context is available to allow erroneous claims to be corrected or argued against.

3. Linked data

3.1 What is linked data?

Linked Data is an approach to exploiting the strengths of the World Wide Web to enable effective large scale discovery, access to and integration of data. The architecture of the web has proven to be extremely scalable and extremely powerful: leading to the enormous collection of information and services now available online. Linked Data is about extending the principles of the web into the domain of structured data.

It incorporates a number of important principles: assign global web-accessible identifiers to real world things like people, places, events; use the mechanisms of the web to provide information about these things via their identifiers; and exploit web links to connect one piece of data to another, to help with discovering new information and comparing or combining it with what you already have.

Sir Tim Berners-Lee described Linked Data via these four principles in his original note on the topic[22]:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs. so that they can discover more things.

To encourage government data owners to make their data available openly and in an accessible way, Berners-Lee developed the 'Five stars of open data', to illustrate the steps an organisation can take, from basic publication through to a fully described machine readable approach.

*	Available on the web (whatever format) but with an open licence, to be Open Data
**	Available as machine-readable structured data (e.g. Excel instead of image scan of a table)
***	As (2) plus non-proprietary format (e.g. CSV instead of Excel)
****	All the above plus: use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
*****	All the above, plus: link your data to other people's data to provide context

[22] http://www.w3.org/DesignIssues/LinkedData

While in many respects, Linked Data is a set of principles, it is most commonly associated with a particular approach to representing data in a machine readable way: namely the 'Resource Description Framework (RDF)[23]. This is a set of specifications and standards, developed and maintained by the W3C and its members.

RDF is based around the concept of representing data using 'triples'. Each triple has a subject, a property and an object. The basic principles of RDF are explained and illustrated in the W3C 'RDF Primer' document[24].

3.2 Why is linked data useful?

Linked data is primarily a data integration technology. Data integration – combining different sources of data together to achieve a particular objective – requires shared identifiers, a shared approach to representing data and its structure and a way of transporting that data.

Because linked data is based around the architecture of the web, and the web is the easiest and most powerful way to distribute open data, linked open data is a particularly powerful concept. Like the web of documents, it is extremely scalable and enables a distributed model of data publication, while providing a way of connecting up and using information from diverse sources.

Note however that linked data can also be applied in a closed 'inside the firewall' environment, where access to the information is controlled. Many large organisations, such as governments or large corporations face significant challenges in exchanging data between different parts of the organisation and linked data is growing in popularity as a tool in tackling this well-known data silo problem.

As explained above, linked data typically uses RDF as a standard way of representing and exchanging data. Few people want to use RDF in its native form, but RDF enables a precise representation of the meaning and structure of data using a standard data model and syntax. By converting different data sources to this lingua franca, they can be shared and combined. It can be converted where required into various other formats that work well with popular software tools, as illustrated in the following diagram.

[23] See http://www.w3.org/RDF/ and http://www.w3.org/standards/techs/rdf for links to the various individual documents
[24] http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

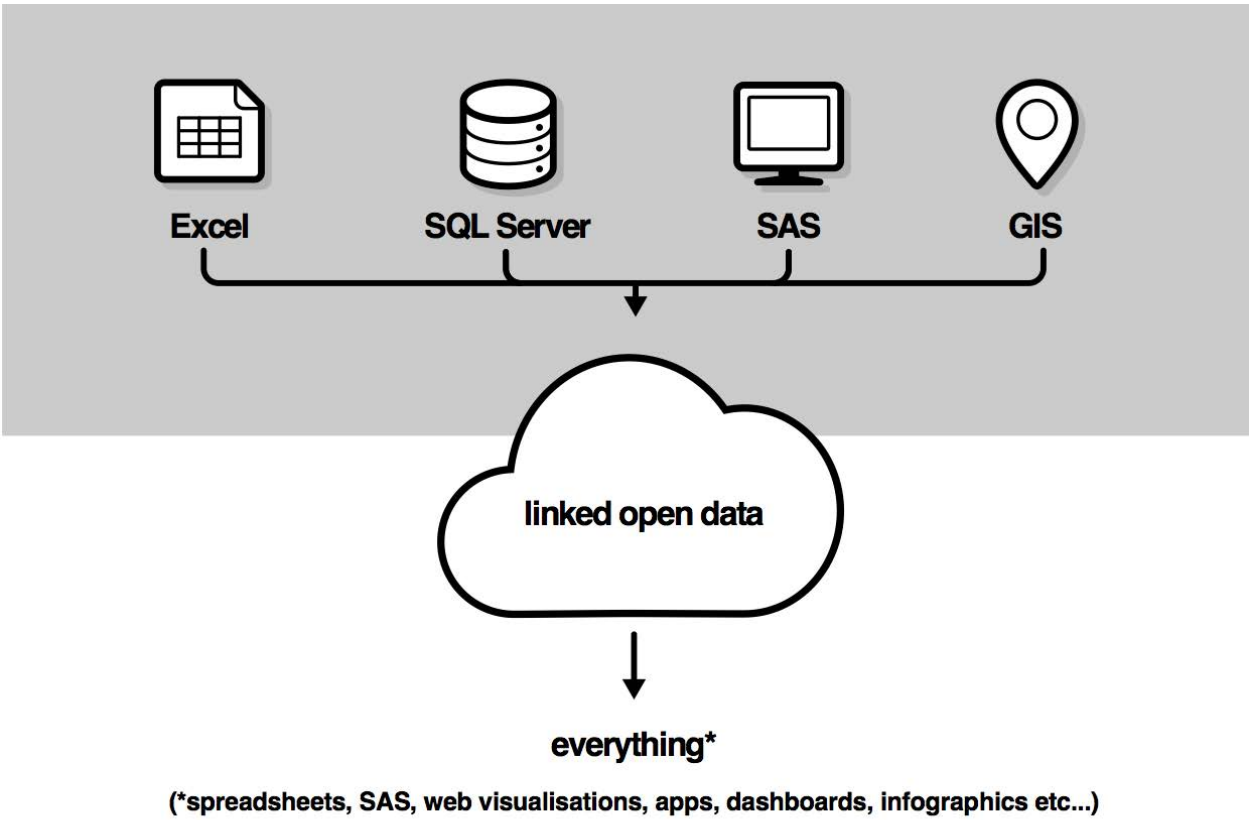

Figure 1 Use of Linked Open Data as an interchange method

3.3 When is linked data a good option?

Linked data is suitable for all kinds of open data publishing, but requires more effort than simply making a spreadsheet or CSV file available for download. Therefore it is worth concentrating, at least in the early stages of an open data strategy, on those datasets which are of greatest interest or value, or which have a role to play as reference data for many other datasets.

In some cases, providing a file download is not a viable approach to making data available and some form of API access is required:

- when data changes frequently
- when the data files are very large
- when you want to select interconnected data from multiple sources

Linked data is a good solution to the first two of these cases and in effect the only feasible solution for satisfying the third case in a web context.

The fact that linked data is a natural approach to maintaining a list of authoritative identifiers and descriptive information about those identifiers makes it an ideal solution to publishing reference data (for example lists of local authorities, or geographical regions) and for controlled lists (e.g. definitive categories for dividing a population into age groups or ethnicities).

3.4 Examples of linked data in practice

The use of Linked Data is becoming more widespread in the public sector in the UK. Some of the most important examples of linked data publication are:

- Ordnance Survey: http://data.ordnancesurvey.co.uk/ Linked data versions of the OS BoundaryLine, CodePoint Open and 50k Gazetteer data products
- Department for Communities and Local Government: http://opendatacommunities.org Statistics on housing, deprivation, planning, local government finance
- Environment Agency: http://environment.data.gov.uk Measurements of bathing water quality and catchment management data
- Office for National Statistics: http://statistics.data.gov.uk Reference information on statistical geography
- Scottish Environmental Protection Agency: http://data.sepa.org.uk/ Reference information on water bodies and catchments
- Land Registry: http://landregistry.data.gov.uk/ Price paid data for housing transactions
- Companies House: http://data.companieshouse.gov.uk/doc/company/SC337356 (for example) Reference information on registered companies
- British Library: http://bnb.data.bl.uk/ Bibliographic metadata on British Library holdings.

4. Requirements for effective use of open data

Earlier in this report we noted that the value of open data arises at the point of use. Therefore to create the greatest return on investment for open data publishing, all aspects of the process should be designed with the use of data in mind.

The ideal situation is that all data which can be open (because it does not contain sensitive personal or security related data) is published openly. Within that long-term objective, priority should be given to releasing data for which there is a known or expected demand. This choice can be made based on the knowledge of the data owners and their discussions with users, or via a more formal user engagement process.

It is then important that potential users of the data know that it exists and is available. This can be tackled in various ways and typically a combination of approaches is best.

- Web search. Many people's first port of call when looking for something is to go to a web search engine. Ensuring that published open data can be found and indexed by search engines will make it easier to find for users
- Catalogue. Maintaining a central list of available government data gives users a central starting point for a data search. To make such a catalogue searchable requires consistent and up to date metadata for each entry. Manually entering information in a catalogue is time consuming and rarely a top priority for data owners, which means that many dataset catalogues suffer from poor quality and incomplete metadata. Finding ways to automate as much of this process as possible is important.
- Social awareness raising. Within various user communities, talking to people (in person, at events or via online forums) is an important part of helping them to know what is available. This can be assisted by technical aspects of a system – for example making it easy to provide a web link to a specific dataset or part of a dataset, so that it can be referred to in an email or on social media.

Once the user finds data that is potentially of interest, it is very important that they are able to make an informed decision on whether it is suitable for their chosen purpose. Datasets need to be associated with good quality metadata that explains:

- Meaning and definitions of terms used in the data
- Where the data came from and how it was processed
- Limitations in how the data can be meaningfully applied
- Quality considerations
- Whether the data will be maintained in the future, or whether it is an experimental, short-term or one-off publication.

This last point is an important part of enabling 'serious' use of data, whether in the private or public sector. For a business to start making use of data in its operations, or to incorporate open data into a product that it sells to its customers, requires up-front investment. Businesses will only be willing to do that if they believe they will have sufficient time to recoup a return on that investment, so need confidence that the data they are using will still be maintained and available for some period in the future. Similar considerations apply to other public sector users of data: incorporating a particular data source into a process or procedure also takes investment.

That doesn't mean that a data owner must commit to maintaining all of their data in perpetuity. There are often good reasons for short term or experimental publishing of some data. The important thing is to communicate clearly to potential users of the data what the publisher's intentions are in this regard.

The 'Open Data Certificate'[25] introduced recently by the Open Data Institute is a useful tool in evaluating whether a dataset is presented in a way that enables and promotes use.

One of the risks often raised around publishing open data is that users will misunderstand the data and use it in ways that are not justified. This risk cannot be completely eradicated, but 'responsible re-use' can be encouraged by ensuring that the data is associated with good quality and thorough metadata and documentation that helps a data user understand what it means, where it came from and possible limitations they should be aware of. This can be embedded in the data itself or linked to from the data.

There is a spectrum of types of users and a range of ways in which they want to make use of open data. To maximise value of data, it is important to serve the needs of all of these users. Some data publishing activities may choose to focus on one audience type, but it is important to acknowledge the range of audiences and consider how their needs will be met.

Figure 2 presents one way of dividing the overall potential audience into categories, based on their objectives in using data and their familiarity with web technologies. For each audience category, it shows the data presentation approaches that are likely to be most appropriate.

An important aspect of linked data is that, although only a small proportion of data users wants to use it directly, it makes it easy to create other formats and other ways of presenting the data to suit the needs of the rest of the data audience. The hard

[25] https://certificates.theodi.org/

work of structuring the data and allowing it to be queried and filtered has been done and that greatly simplifies the process of generating the other forms of presentation.

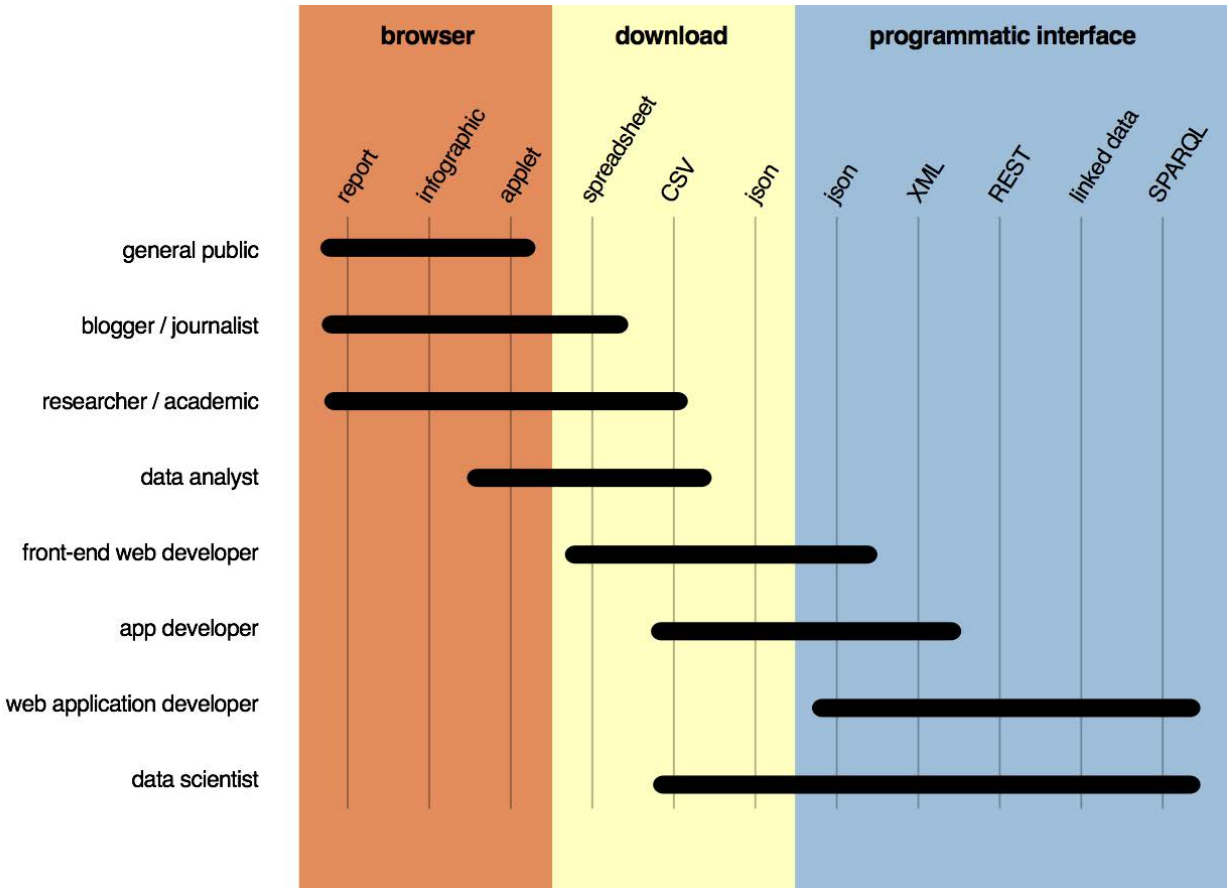

Figure 2 Publishing open data for multiple audiences

The general public, bloggers, journalists and researchers most often want to view data in their browser in the form of reports, infographics or simple apps. In some cases they may want to download the underlying data for a closer look in their preferred software tools.

For researchers and data analysts, the ability to download data for further processing is a higher priority. The choice of format for downloads depends on the type of data and its likely uses (and the diagram does not attempt to list all possibilities) but CSV or spreadsheet formats are the most popular.

Web developers generally want to take a selection of the available data and to incorporate that in their own application. They are often working to create a tailored presentation of information for a particular audience. Sometimes a developer will download data and manage it in their own database, but in most cases some form of programmatic interface to the data is preferred.

A subset of developers will use linked data or SPARQL directly, often to act as an intermediary serving the needs of a user group further up the spectrum.

Most users, even the most technical ones, will make their first contact with a published collection of data by reviewing web pages and visualisations to get a feel for what data is available, before digging deeper to see how the data is structured and how they can access it in a machine readable way.

All users of the data are therefore well served by thoughtful design. Once a user finds a website that presents data to them, whether they continue working with it, or abandon it and look elsewhere, is heavily influenced by their initial experience of using the site. Careful consideration of the user experience of the site is important: this incorporates information architecture, navigation processes, page layout and speed of response of the site for common tasks.

5. Pilot of linked open data publishing

The open data consultancy study included establishing Open Data Scotland (http://www.opendatascotland.org). This is a pilot data publishing website intended to demonstrate the possibilities of the linked open data approach and to investigate the feasibility and implications of applying this approach more widely to publishing Scottish Government Statistics.

The site incorporates:

- A selection of existing datasets published as linked data: the Scottish Index of Multiple Deprivation (SIMD) for 2004, 2006, 2009 and 2012; school establishments and associated data on school pupil and school leaver numbers; supporting geographical datasets including council areas, data zones, intermediate zones and postcodes
- A showcase website for SIMD, including visualisations and data navigation features, built on top of the linked data SPARQL endpoint (http://simd.opendatascotland.org/)
- A tutorial on using open data in a web context, based on the schools data incorporated in the site (http://schools.opendatascotland.org/)
- A 'technical' section of the site, aimed primarily at developers and data specialists, where the datasets can be explored, and accessed in machine readable forms via a range of Application Programming Interfaces (APIs). (http://data.opendatascotland.org/)

Our main objectives with the pilot were to consider the range of users of Scottish Government analytical services data in general, and of the selected datasets in particular. A particular emphasis was placed on potential new users of the data. Exploiting the possibilities offered by the linked data approach, we have experimented with new data presentation approaches to meet their needs.

5.1 Details of the approach

5.1.1 Overview

The starting point of the pilot was the selection of datasets to be incorporated. It was agreed to use a small representative selection of datasets that we could work with in depth.

The Scottish Index of Multiple Deprivation was chosen as it is already a widely used dataset with many applications. We chose to include datasets about schools in the pilot, because education data is another area with broad interest, and using school-based data required us to tackle data with a different structure to the majority of SNS datasets, which are mainly statistical data organised by council area or by data zone.

In addition we needed to include geographical reference data: we created linked data datasets of council areas, intermediate zones and data zones, the hierarchical relationships between them and the connection from each area to a representation of its boundary. We also included a dataset of postcodes, as these are popular and widely used as a geographical reference point. The postcode data was obtained from the Ordnance Survey and the Office for National Statistics. It includes the coordinates of postcode centroids and a link from each postcode to various geographical region that contain that postcode.

Via the pilot, we wanted to explore and demonstrate the following points:

- Provide data in linked data form to enable queryable access to a collection of datasets
- To use the strengths of the web in presenting the data, by ensuring that all data collections and views have a persistent URL that can be linked to
- To show that it is straightforward to create dynamic visualisations that draw data live from the underlying linked data database
- To connect user-friendly presentations with access to the underlying data, so that others can create their own visualisations.
- To provide learning materials and data access that encourage others to create their own uses of the data

In doing this, we aimed to meet the needs of each of the audience types discussed in the chapter 'Requirements for Effective Use of Open Data'.

5.1.2 Representing the data as linked data

The starting point of the work was to convert data from its original formats to '5 star' open data in RDF format. We followed the approach set out in the Linked Data Cookbook[26], as documented by the World Wide Web Consortium Government Linked Data Working Group. The Cookbook approach is summarised as follows:

> *"The 7 Best Practices for Producing Linked Data*
>
> 1. *Model the Data*
> 2. *Name things with URIs*
> 3. *Re-use vocabularies whenever possible*
> 4. *Publish human and machine readable descriptions*
> 5. *Convert data to RDF*
> 6. *Specify an appropriate license*
> 7. *Host the Linked Data Set Publicly and Announce it!"*

[26] http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

Of these, steps 1, 2, 3 and 5 are the most onerous in terms of effort and required knowledge. We will describe each of the steps in more detail:

1. Modelling the data is essentially a process of understanding the structure of the data and deciding how to represent it in the triple based structure used in RDF. It involves identifying the most important entities in the data, what attributes these entities have, and how the entities are related to each other and to external entities that we might refer to but not describe in detail.

   In carrying out this process, we drew on experience of common data structures and good ways of representing these in RDF. A very useful resource in this process is the online book 'Linked Data Patterns'[27] by Dodds and Davis, which documents a large number of frequently encountered data structures with notes on how to represent them as RDF.

2. In RDF all resources are assigned a URI (Uniform Resource Identifier). Therefore having decided what the important entities are in the data, we need to decide how to assign URIs to them. The most important decision at this point is to decide whether the entities of interest already have a maintained URI from a trusted source. If that is the case, it is generally good practice to use that identifier. In our case, we chose to use ONS identifiers for geographical areas and data.gov.uk identifiers for time intervals. This is an aspect of the 'linking' part of linked data, which makes it easy for users of the data to combine data from different sources.

   If no suitable identifier already exists, we have to create our own. As explained in the Linked Data Cookbook, this needs to be done in a web domain that we control – in this case http://opendatascotland.org. In designing the format of these URIs, we followed the advice in the URI Patterns[28] document developed by the UK Linked Data Working Group.

3. Step 2 relates to identifying entities in the system. We also need to assign URIs to properties and classes that describe the interrelations of the main entities. These properties and classes are often referred to as a vocabulary or ontology. As with identifiers for entities, a first step is to consider whether a suitable vocabulary already exists in a trusted well-maintained form. If it does, it is good practice to re-use it as this makes it easier for users to understand the data and relate it to other data sources. When using an existing vocabulary, it is important to consider carefully if its documented meaning is a good match for our context. Existing terms should only be used if we accept that their documented meaning is the meaning we want to express. The

[27] http://patterns.dataincubator.org/book/
[28] https://docs.google.com/a/swirrl.com/document/d/1Id8GSMAgiWWOaKsn1TUqXPgZ17tyoKp2oKqM9UvDnjE/edit#

Linked Data Cookbook lists some of the most commonly used vocabularies[29]. A useful resource for discovering existing vocabularies is the 'Linked Open Vocabularies' site[30] developed by the DataLift project.

A particularly important vocabulary for Open Data Scotland and for many statistical data applications is the RDF Data Cube[31], currently going through the W3C standardisation process.

4. A principle of linked data is that if you look up the identifier for an entity (often called a 'resource') then the system should return a description of that entity. Best practice is to provide both human readable descriptions in HTML form as well as machine readable descriptions, possibly in a variety of formats. The 'content negotiation' methods built in to the standard HTTP protocol[32] are used to specify which format a user wants. This linked data resource look-up approach is a standard feature of Swirrl's PublishMyData platform that we used as the basis of the pilot.

5. Once we have decided the structure of the data and the system of identifiers we want to use, then we can go through the mechanics of converting the data into RDF. Our starting point for this project was a set of spreadsheets and CSV files. Our approach was to write simple scripts in the Ruby programming language, drawing on established code libraries that handle much of the details. These scripts read in the source data and then work through each data point and output the RDF representation of them.

   This kind of data processing script is reasonably straightforward to create for someone with modest programming skills. Supporting libraries for working with RDF exist in most programming languages.

   Some graphical user interface driven tools exist for converting tabular data to RDF, though the process of configuring these tools is often just as complex as simple programming.

   Swirrl, through the EU funded 'OpenCube'[33] project is working on user-friendly tools to support this data conversion process.

6. It is important that users of the system know what they are allowed to do with the data. Therefore a data publisher should always specify a specific licence

[29] http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook#Step_3_Re-use_Vocabularies_Whenever_Possible
[30] http://lov.okfn.org/dataset/lov/
[31] http://www.w3.org/TR/vocab-data-cube/
[32] http://en.wikipedia.org/wiki/Content_negotiation
[33] http://www.opencube-project.eu/

for any data they publish. In the case of Open Data Scotland, all Scottish Government owned data is released under the Open Government Licence[34]. The system also incorporates some data that belongs to the Ordnance Survey and in these cases the data is made available under the Ordnance Survey Open Data Licence[35].

7. The datasets we have created are hosted at the site http://www.opendatascotland.org that consists of user-friendly navigation and visualisation pages as well as direct access to the underlying linked data via http://data.opendatascotland.org

5.1.3 Structure of the system

The Open Data Scotland site is built on top of Swirrl's PublishMyData[36] linked data publishing platform. The basic structure of the system is illustrated in Figure 3. The core of the system is the triple store, which uses the open source Apache Jena system and holds the RDF data used by the site. The data in the triple store is accessed via SPARQL queries and used to create the dataset navigation and browsing facilities at http://data.opendatascotland.org and to extract data required for the visualisations at http://simd.opendatascotland.org. The geographical visualisations also make use of a set of geographical boundary files which are stored on the server in TopoJSON format. These TopoJSON files were created from original data in ESRI Shapefile format, provided by the Scottish Government. TopoJSON was chosen as a working format because it is compact and easy to use in our chosen visualisation libraries (see next section).

As well as being used internally to generate web pages and graphics, the SPARQL endpoint is also directly available to external users who want to run their own queries against the database.

The system incorporates an administration interface, used for creating and updating datasets and their metadata as well as gathering analytics on system use.

The triple store, administration interface, SPARQL endpoint and Linked Data browsing are all parts of the standard PublishMyData platform. The web pages and visualisations available at http://www.opendatascotland.org, http://simd.opendatascotland.org and http://schools.opendatascotland.org/ were designed and created specifically for this project.

[34] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
[35] http://www.ordnancesurvey.co.uk/docs/licences/os-opendata-licence.pdf
[36] http://www.swirrl.com/publishmydata

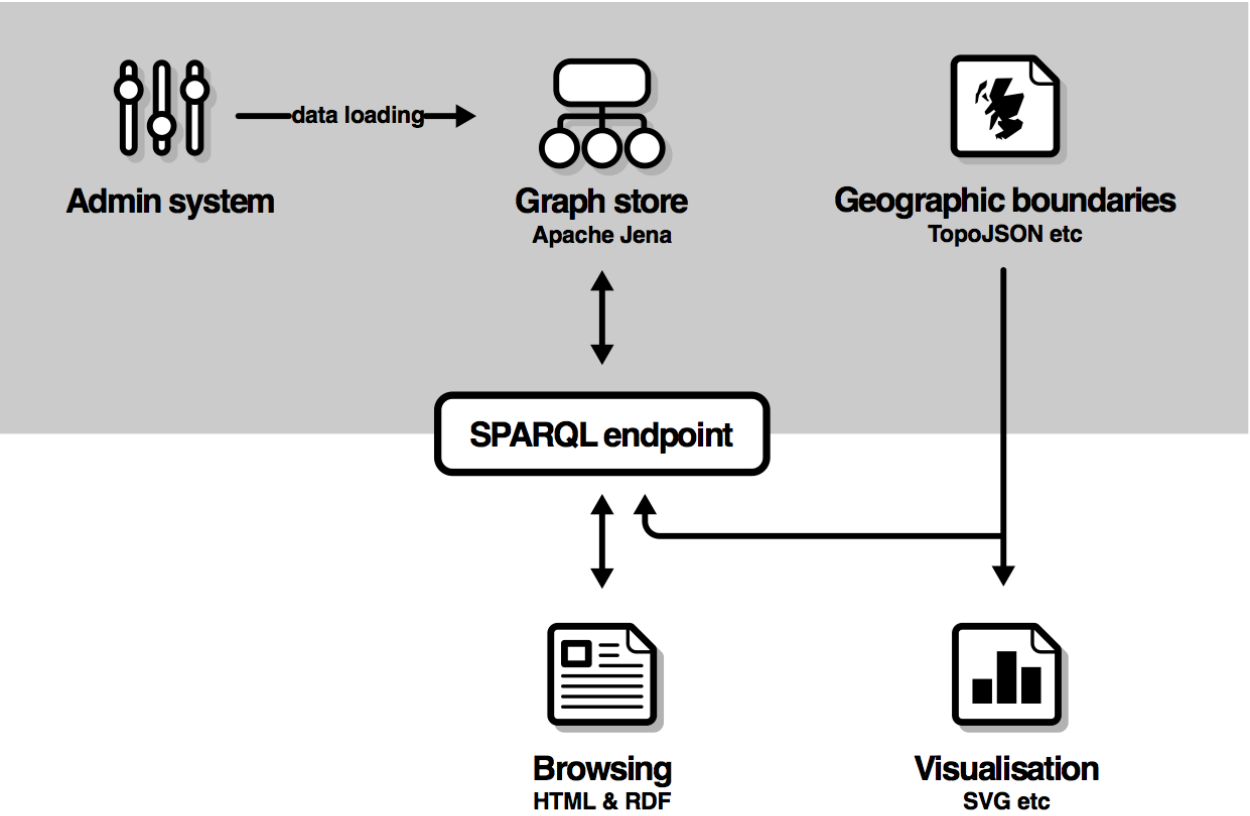

Figure 3 High level system structure

5.1.4 Approach to creating visualisations

The visualisations in the site were created using HTML, CSS and Javascript. The main Javascript library used was D3 ('Data Driven Documents')[37]. The visualisations obtain their data by running SPARQL queries on the PublishMyData SPARQL endpoint.

The architecture of these visualisations is such that they could be hosted on any website – they use the publicly available query endpoint to get their data and they are created with open source libraries.

It is an illustration of how the Linked Data approach enables not only the data owner to create a data portal, but allows third parties to build rich views of the data to suit their own purposes.

[37] http://d3js.org/

5.2 How could this approach be applied more widely?

The pilot incorporates only a small sample of the data available from the Scottish Neighbourhood Statistics site, together with other related education and geographical data.

What would be involved in applying this approach to the entire Scottish Neighbourhood Statistics site?

The overall SNS contains around 1000 indicators, many available for a range of time periods and in some cases at more than one level of geography.

The basic pattern applied in the pilot to representing the SIMD as linked data is well suited for extension to the whole of SNS. Because of the large number of indicators, it will be necessary to apply a reasonably high degree of automation to the process. This is certainly feasible and the process will benefit from the high degree of structural similarity between the different datasets.

The SNS data all fits well into the RDF Data Cube model. Each dataset will have a geography dimension and a time dimension and often has one or more further dimensions such as age range or gender.

Where an indicator is available for multiple time periods, then in general all time periods should be included in the same Data Cube dataset (as long as other dimensions and the method of indicator calculation remain the same across all time periods).

In many cases, several indicators should be combined into a single Data Cube dataset, where the same quantity is being reported but for several values of a dimension such as age-range. In SNS for example, there are separate employment level indicators for age ranges 16-24, 25-34, 35-49, 50-64, as well as a 'total age range of 16+. These can be combined into a single employment levels dataset, with a dimension of age range.

The total size of a linked data version of SNS is therefore likely to run to several hundred datasets. The number of triples in the system will be dominated by those indicators that are provided for the smallest (and hence most numerous) geographical areas – i.e. at data zone level. (A dataset at local authority level will be around 200 times smaller than a data zone dataset).

A data cube dataset typically includes around 5-8 triples per data point. Therefore for a data-zone based dataset, we can expect an average of around 40,000 triples per indicator per time period. If there are say 500 indicators at data zone level with

an average of 5 time periods each, then we will have around 100,000,000 triples in the data store. This is well within the standard capabilities of established triple stores and so would not present significant technical challenges.

To establish the patterns in the data that can be automated will require a review of the indicators, dimensions and values to develop a series of URI patterns that can be used consistently across the SNS data. This will involve a decision on the top level domain that the data should be contained in and the URI patterns should be designed taking into account possible future Scottish Government linked data publishing on other topics.

This analysis should include identifying where the same dimensions or dimension values appear in connection with multiple indicators and so can be re-used. For example the same set of age ranges is used for the Employment Level indicators and Employment Rate indicators (amongst others). A particular age range should be described using the same URI wherever it appears.

The pilot has produced geographical datasets for council areas, intermediate zones, data zones and postcodes. To cover the whole SNS will require a number of other geographical datasets to be established: health boards, wards, community health partnerships, parliamentary constituencies and community regeneration areas.

An important aspect of the data creation is providing metadata and documentation about the data. SNS already includes systematic metadata about indicators. As this is in a consistent form, it could be extracted from the original system and processed into a linked data form. It would be useful to consider how this could be improved and extended in future. There are opportunities to create or link to existing richer descriptions of the background of how data collection and data processing has been carried out. The administration system for managers of a new linked data SNS will need to incorporate a user interface for managing and editing metadata.

A system will be required for managing the data in the system, principally for statisticians to add updates as new data becomes available, but also to allow correction of errors, improvements to metadata and so on. This should be designed in consultation with the statisticians responsible for maintaining the data. Such a system could be designed to work in a similar way to the existing SNS data entry system, to minimise unnecessary changes in working patterns.

It would be possible, and highly desirable, to provide an API for adding and updating data (in addition to a user-interface driven process). This creates possibilities for greater automation of SNS data maintenance in future.

An important aspect to consider is how to design the site navigation to enable a user to find and access the data they are interested in. The pilot site does not tackle this

navigation question in depth as it incorporates only a small number of datasets, although there are elements in the navigation design of the pilot that could form a useful part of a broader system. Generally speaking, we anticipate the need to allow data navigation both by theme and by geography. The system should allow both browsing and google-style search.

It is not necessary and probably not practical to present all datasets in the SNS in the rich way that we have implemented for the SIMD data in the pilot. The majority of datasets could have a simpler, but still user friendly, presentation. However, the approach taken to mapping the SIMD in the pilot could be easily extended to other datasets and other geographies.

Given the richness of the data in SNS, there are many other opportunities for developing interesting visual ways of exploring and navigating the data and if the approaches demonstrated in the pilot were to be extended to the whole of SNS, it would be useful to investigate ideas in this area.

5.3 Assessing success of the pilot

At the time of writing of this report, Open Data Scotland has just gone live and become accessible to the public. To help decide whether to apply this approach more widely, it will be important to gather feedback on how users of the data react to the site: to find which aspects of it they find useful, which aspects might need modification or improvement and how users regard Open Data Scotland in comparison to existing data sources such as http://www.sns.gov.uk or the various spreadsheets that can be downloaded from http://www.scotland.gov.uk.

The user engagement process has therefore only just begun and it is too soon to report on user views. It is important that sufficient effort is devoted to identifying and communicating with users and potential users of the site.

User engagement will involve informing existing and potential users of SNS of the existence of the site, assisting them where necessary in getting started with using Open Data Scotland and gathering their feedback on what they like or don't like about the site.

Overall use of the system will be monitored using Google Analytics and PublishMyData API Analytics and regular reports provided to the Scottish Government.

6. Approach to internal data management

The linked open data approach described above is a good solution to many of the problems of data discovery and data sharing in an open web environment. Many of the same issues are encountered inside any large organisation and so it is interesting to consider whether open data technology is also applicable to internal data management and data exchange.

Many tasks of researchers and analysts within the government require accessing and combining data from multiple sources. This requires finding what is available, understanding how it was produced, what limitations and constraints the data might have, then processing the data into the required form.

Linked data is essentially a data integration technology, well suited to distributed and diverse collections of data, such as those created and used within the Scottish Government. It is becoming more popular for enterprise data integration and business intelligence applications.

In this chapter we consider the implications of applying the linked data approach to internal data management.

The Analytical Leadership Group of the Scottish Government recently considered a paper on 'Strategic Data Management' setting out the main requirements for management and exchange of data within the analytical services divisions.

This document identifies the main objectives of a new solution to strategic data management as:

1. *Provide visibility of analytical datasets held across SG to all analysts*
2. *Enable access to data from the widest possible range of analytical tools*
3. *Ensure the security and integrity of data*
4. *Provide metadata on the datasets we hold and the data within them*
5. *Empower analysts to manage their data effectively within a clear framework*
6. *Automate as much as possible*

We will consider how a linked data approach would relate to each of these main objectives.

1. The issues around discovery of open data are very relevant to this issue. Web technologies in general (corporate intranets for example) are highly applicable to 'inside the firewall' applications where there is a medium to large number of users, particularly when they are distributed across several locations.

Our discussions with government analysts indicated that personal contacts –

‘knowing who to ask’ – is often the first port of call when trying to find information from outside their own domain. This will always be a useful component of data exchange, but a more systematic approach can help people to discover data more quickly and more reliably. Picking up the phone to the appropriate person may still be necessary or useful for detailed questions, but a better approach to dataset cataloguing and dataset metadata will reduce the workload on individuals and will allow analysts to find data that they may otherwise have missed.

Creating a simple standardised set of dataset metadata and tools to support the creation and maintenance of that metadata (automated where possible) allows the creation of browsing and search tools that will improve data discovery within Analytical Services.

2. Commonly used tools include SAS, Excel, SPSS, SQL Server. There has been a significant investment in the use of these tools, in terms of staff expertise and development of customised workflows and other software. Any new solution to data management must allow these tools to continue to be used, and to be flexible enough to allow the introduction of new analysis tools in future.

If data is to be held as linked data in the underlying data management system, this requires the structure and meaning of the data to be made explicit. This requires effort, but generally only needs to be done once for each dataset or group of related datasets and provides benefits in repeated more effective uses of that data.

There are two aspects to consider: how data in linked data form is created and updated; and how linked data is accessed via existing tools.

The majority of the data analysis tools in use are based on essentially tabular data. Statistical linked data consists mainly of n-dimensional data cubes, supported by reference data in list or tree structures. By choosing 2 dimensions at a time, data cube data can be converted easily into a table or series of tables. Some format conversion work will be required to extract data in a form that can be read into tools like Excel and SAS but this is relatively straightforward.

The task of creating and editing linked data from tabular data tools is more difficult and is the subject of active research and development, for example in the EU funded OpenCube project. It requires a mapping of data from a rows and columns format to the triple-based representation of linked data, involving consideration of the ‘Linked Data Cookbook’ steps described in Section 5.1.2 of this report. A series of data mappings will need to be developed and maintained, and used to support the integration of data analysis tools with the underlying data management store. This is the most technically complex aspect of using a linked data approach to the data management system, but is also the aspect that brings significant benefits as it enables a rich integration of data from different sources.

3. Data held in the data management system will be subject to access restrictions as some of it will include personal data or other sensitive material. Therefore, not only will the system as a whole need to be secure from unauthorised access, it must be possible to manage access to datasets within the system. Using currently available linked data technology, there are a number of possible approaches to this issue. The challenge here is that one of main strengths of linked data, the ability to use SPARQL to query across a number of interlinked datasets, is a feature that must be carefully controlled if access to specific parts of the data are to be controlled.

There are 3 main levels of data organisation in a triple store: the triple, the named graph and the database (sometimes referred to as the 'dataset' in SPARQL documentation, but this is a different use of the word 'dataset' to the rest of this report, so we will avoid that term). A typical triple store platform can host many databases simultaneously. In most triple store implementations, a SPARQL query can access any data in a single database. Some stores may have extensions that allow finer grained access control for SPARQL queries, but this is not standard, so it would restrict technology options to rely on that. Access to individual resources or individual named graphs ('buckets' of data within a database) can be controlled in a user interface layer for general data presentation and data browsing, but is hard to manage reliably if SPARQL query access is offered.

A reasonable approach may be to manage access at the database level, which allows a number of reliable security technology options. This means that data should be grouped into databases based on consistent access requirements. This is a similar concept to controlling access to SAS data at the 'library' level.

4. Providing consistent metadata on datasets and their contents is an important requirement for making the data easy to find and for an analyst to decide if it is suitable for their purpose. This objective has much in common with objective 1 in terms of the technical requirements for meeting it.

5. To empower analysts to manage data will require providing good management tools for reviewing datasets in a database, controlling who has access to it, creating and maintaining metadata, managing versioning and so on. Due attention will need to be paid to a user friendly interface to support these functions.

6. The linked data approach incorporates open standards for interfaces and APIs. Consistent APIs and consistent standards for data representation means that a high degree of automation is possible. In addition to the standard facilities offered by linked data systems, some form of workflow management will be required. There are many existing options for doing this and the choice of technology for workflow management would depend on the complexity of the processes to be automated and

the frequency with which they are modified.

Using linked data internal data management will significantly reduce the additional effort required for open data publishing as the hard work of analyzing and representing explicitly the data structure and meaning will already have been done.

Use of linked data for internal data management, with access control and versioning, is less well established than use of linked data for open data publishing. To learn in more detail about the opportunities and possible pitfalls, we suggest carrying out and reviewing a small scale experimental implementation of the approach.

7. Enabling effective uptake of linked open data across the Scottish public sector

Earlier in this report we have described the potential benefits that can be derived from using open data to make public sector information more accessible. We have described the process and infrastructure involved in creating the Open Data Scotland pilot and discussed a possible approach to applying open data technologies to the needs of internal data management.

If, after assessing the outcomes of this study, the Scottish Government wishes to apply this approach more widely, what steps must it take to enable successful large scale implementation of the open data approach?

Moving to a situation where all or most of the government's non-personal data is available as linked open data is a significant task and can't happen overnight. Any 'big bang' approach is likely to fail. This chapter sets out the most important supporting activities that need to be put in place and a high level vision of how such a transition might work.

Broadly speaking, the necessary activities and changes required for a large scale implementation of linked open data publishing can be grouped into four main categories:

- Standards
- Knowledge and skills
- Tools
- Culture

We will address each of these in turn.

7.1 Standards

By 'standards', we are referring to the set of technical specifications and procedures that the Scottish Government should mandate, to ensure that open data production is sufficiently consistent. This will enable data from different parts of government to be interoperable, will give users of the data a consistent and predictable experience when using government data and ensure high quality.

Choosing how much to standardise is a matter of judgement. Being too rigid leads to a central monolithic approach that is too slow moving and too inflexible to succeed. Being too flexible means that different publishing initiatives are disjointed and some may be of insufficient quality.

In the view of the authors, the government should aim to enable a new ecosystem of data publishing and consumption. This ecosystem should be based around consistent patterns, but allow a distributed approach. Different parts of the public sector can choose their own software solutions, choose whether to do things in house or to use external suppliers, have a choice of external suppliers, decide to do things independently or to share facilities with other parts of the public sector and so on: as long as they act within a framework that ensures sufficient consistency and interoperability. Setting and applying common standards avoids vendor lock-in and promotes competition amongst suppliers. The distributed decentralised approach spreads the effort, ensures that people who understand the data are directly involved in the process of distributing it and encourages innovation in this developing area.

By definition, adopting a linked data approach brings with it the obligation to meet certain non-proprietary technical standards established by the IETF[38] and W3C, around use of HTTP as a data transfer protocol, RDF as a data representation format, and provision of specific HTTP-based Application Programming Interfaces. This already leads to a high level of interoperability.

The Scottish Government will need to agree an additional layer of standards, to sit on top of these fundamentals.

These should cover:

- Design of URI patterns
- A list of standard vocabularies for commonly occurring types of information
- Sets of URIs and associated reference information for 'core reference data' – the kinds of data that get referred to in many other datasets and are a key point of interconnection for different data. This includes identifiers for geographical areas, time intervals, important government assets, government departments and so on.
- Guidelines for how public sector publishers should go about extending the core standards to meet their own more specific needs.
- Metadata standards for describing datasets and the processes that have led to their creation. This could follow or be inspired by the requirements of the Open Data Certificate from ODI, as well as W3C metadata schemes such as DCAT[39] and VOID[40].

At the next level of detail, individual departments or groups may develop their own more specific standards for commonly occurring concepts, for example an ontology to represent measures of economic activity, or a concept scheme for age ranges that can be used across all SNS indicators that have a breakdown by age.

[38] http://www.ietf.org/
[39] http://www.w3.org/TR/vocab-dcat/
[40] http://www.w3.org/TR/void/

7.2 Knowledge and Skills

The approach we have described here requires some new approaches that are new to most statisticians and other public sector staff working with data. It is important that the government develops a base level of awareness and knowledge of open data and linked data amongst those who work regularly with data; and also a centre of expertise within government that can provide advice to others and support learning.

Although day to day activities can be supported by appropriate tools and assistance from external experts can be used where required, it is important that the overall view of open data policy is internally led. Sufficient knowledge of the technologies and standards must exist within the government to support sound decisions, whether on operational details or during procurement processes.

The range of relevant skills includes:

- Understanding of the RDF approach to data representation
- Data modelling for linked data and familiarity with common practices for URI and vocabulary design
- Programming techniques for data transformations
- Programming techniques for presenting data on the web

Developing a skill base in this area will take time and planning, but the essential background and education needed is already in place amongst many staff in Analytical Services, Information Services and elsewhere in the Scottish Government.

7.3 Tools

There is already an established and maturing market in open source and proprietary software for triple stores to support the underlying storage and querying of linked data. Presentation of open and linked data on the web tends to rely on commonly available web development and visualisation tools.

The main area where further tool development is required is in supporting the process of transforming data from its native form into RDF – to support steps 2, 3 and 5 of the Linked Data Cookbook approach described in Section 5.1.2.

A variety of software tools exist to support this process but none is very mature and more development is required. In the area of statistical data, many datasets are best represented using the RDF Data Cube approach – this common structure allows

narrowing the problem to be tackled in software tool development and so is a natural area to concentrate initial effort.

This is a developing area and sharing challenges, experience and expertise with other public sector open data publishers is to be recommended.

7.4 Culture

The experience of the authors during this open data consultancy project has been that the Scottish Government staff have been enthusiastic to investigate new possibilities that might allow them to improve their service to the public. There is already a culture that understands the growing importance of and opportunities arising from applying digital technology to public services.

Implementing any significantly new approach in a large organisation will always face challenges however. There is the unavoidable inertia that comes with any large organisation with a long history (whether private or public sector). The public sector emphasises reliability and trustworthiness and this naturally leads to an element of risk aversion. When dealing with publishing data that is derived from personal or sensitive data, it is of course very important that new approaches maintain the existing care around proper handling of personal data.

A gradual introduction of new technology, with ongoing reviews of benefits and engagement with the users, allows the culture to evolve alongside the use of technology.

Application of the 'Open Data Engagement'[41] approach proposed by Tim Davies and others assists in developing a new mindset where a primary purpose of the main curators of a collection of data becomes to communicate this effectively to others who may want to use it.

To make this new data ecosystem work will require establishing a culture and goals, where the success of a data curator is measured by how others use their data. The value of open data is created when someone puts it to good use: choices of technology and design of processes should be made with that in mind.

[41] http://www.opendataimpacts.net/engagement/

ISBN: 978-1-78412-190-7 (web only)

The Scottish Government
St Andrew's House
Edinburgh
EH1 3DG

Produced for the Scottish Government by APS Group Scotland
DPPAS19745 (12/13)

Published by the Scottish Government, December 2013