SHAKESPEARE REVIEW

An Independent Review of Public Sector Information

MAY 2013

Contents

The Foundation of this Review...

In October 2012, I was invited by government to lead an independent review of Public Sector Information (PSI) to explore the growth opportunities of, and how to widen access to, the wealth of information held by the public sector. I have based my Review on extensive consultation with numerous stakeholders, and edits by many experts. I would like to thank the following for their contributions:

The economic study by Deloitte, commissioned for this Review and published alongside it;

Attendees from a variety of breakfast seminars and other meetings, including: Neville Merritt (Advanced Business Solutions), David Rhind (APPSI), Cathy Emmas (AstraZeneca), David Dinsdale (Atos), Oli Bartlett (BBC), Nick Pickles (Big Brother Watch), Dr Sonia Sousa (Big Innovation Centre), Adrian Brown (Boston Consulting Group), Mark Langdale (BT), Tom Loia (Bull Computing), Paul Maltby (Cabinet Office), Ed Parkes (Cabinet Office), David Doyle (CapGemini), David Behan (Care Quality Commission), Matthew Trimming (Cognizant Technology Solutions), Alex Coley (DEFRA), George McMeekin (Dell), Steve Dauncey (Dell), Demographic User Group, Peter Knight (Department of Health), Caroline Walton (Dollar Financial UK Limited), Pete Sinden (Dr. Foster Intelligence), Filomena La Porta (EDF Energy), James Nolan (EE), Dave Reynolds (Epimorphics), Trevor Fenwick (Euromonitor International), Steve Harris (Experian), Adam Swash (Experian), Paul Maylon (Experian), Liam Maxwell (GDS), Sarah Hunter (Google), Edwina Dunn (HD Ventures), Clive Humby (HD Ventures), James Johns (Hewlett Packard), Mike Hawkins (HMRC), Dixit Shah (IBM UK), Craig Summers (IBM UK), Professor David Edwards (Imperial University and KCL), Saul Klein (Index Ventures), James Alexander (Intellect), Theodora Kalessi (Intellect), Richard Copland (Logica), Richard Stephens (LORS), Ben Goldacre (LSHTM), Steven Bond (Marks and Spencer), Francine Bennett (Mastodon C), Jamie Cattell (McKinsey MDI), Tim Trailor (MEM Consumer Finance Ltd), Victor Henning (Mendeley), Marc Tellentire (Methods), John Parkinson (MHRA), Ricky Spencer (Mobrey Ltd), Hugo Boylan (Newgrove), Guy Herbert (NO2ID), Rohan Silva (Number 10), John Sheridan (Office of Public Sector Information), Simon English (Open Text), Roger Lee (Oracle), Chris Royles (Oracle), Greg Hadfield (Organiser of Open Data Cities), Roger Goss (Patient Concern), Dr Paul Hodgkin (Patient Opinion), Colin Campbell (PDMS), Alex Warents (Pinsent Masons), Toby Stevens (pixIDust Limited), Mike Thacker (Porism), Mike Sweeney (Post Office), Prof. John Domingue (Knowledge Media Institute at the Open University), Mark Tuley (PROLINX), Richard Barborosa (Red Hat), Dominic Cheetham (RedKite), Thomas Cawston (Reform), Julia Greenfield (SAP), Daniel Hulme (Satalia), David Pegg (SCISYS), Peter Clarke (Severne Ltd), Feargal Hogan (Shipping Guides Ltd), Nik Mughal (Software AG), Hadley Beeman (Technology Strategy Board), Nick Riley (The Money Shop), Prof. Philip Treleaven (UCL), Anwen Robinson (Unit4), Professor Ian Horrocks (University of Oxford), Dr Geoff Nicholls (University of Oxford), Professor Steven Roberts (University of Oxford), Adrian Hawkes (Valpak Ltd), Andrew Clough (VisionWare), Chris Handley (Vodafone), Neil Smith (Wilmington Group plc), Polly Avgherinos (Wilmington Group plc).

The Open Data Institute (ODI) for organising and hosting the Open Data Market Makers event and to the individuals who attended for their contributions. And the ODI's chief luminaries, Nigel Shadbolt and Gavin Starks.

Special thanks for help with drafting and revising, and deliberations among the Data

Strategy Board including: Dr Lisbet Rausing (Imperial College), Professor Peter Baldwin (University of California), Lord Redesdale (Energy Managers Association), John Dodds (BIS), Tim Kelsey (NHS England), Steve Thomas (Experian Plc), Sir Ian Magee (GICG), Professor Mark McGurk (GKT Dental Institute), Sir Mark Walport (Government Chief Scientific Adviser), Chris Graham (Information Commissioner), Michael Coughlin (LGA), Jane Frost (Market Research Society), Carol Tullo (National Archives), Nigel Shadbolt (ODI), Heather Savory (Open Data User Group), Rufus Pollock (Open Knowledge Foundation), Claudia Arney (PGD), Chris Yiu (Policy Exchange), Nick Baldwin (PWSCG), Bill Roberts (Swirrl IT Ltd), Professor Terry Lyons (University of Oxford), Peter Elias (Warwick University), David Dobrin (YouGov), Tilly Heald (YouGov)

Respondents to the two waves of surveys, each with two different samples, a) respondents to an open platform promoted to data/policy activists and experts, and b) a sample of the general public derived from a nationally representative YouGov sample; with both samples giving strong support to all the recommendations in the versions they were presented. The findings are available at http://research.yougov.co.uk/

Experts from across government, and my excellent team at BIS, responsible for advising, drafting, organising and keeping the review on track: Steve Brown, Paul Driver, Liz Farmer, Angela Latta, Kanishka Narayan, Matthew Pearce, Frances Pottier, Jane Simmonds, Jennifer Wallace and Ed Woolley.

The recommendations within the Review have been informed by many contributions, some from people with wide generic knowledge of the field, many with specialist expertise. I am indebted to them all for their support. However, the content of this Review and its recommendations remain my responsibility and mine alone.

Foreword

The Revolution, Phase 2: How Britain Can Be The Winner

The digital revolution is entering a new phase. First it was about connectivity, bringing together people, organisations and businesses in new ways that hugely increased communications, access to information and the efficiency of operations. America was clearly the winner, enabled by a large single market, heavy investment in the required basic science and technical application, as well as an innovative and entrepreneurial culture. Think of Google, Ebay, Facebook, Amazon, PayPal, Yahoo, Microsoft, Twitter, Apple – the companies through which our daily lives are run, all headquartered on the West Coast of America.

Phase 2 sees an equivalent leap, this time in the capacity to process and learn from data. Is that exciting? It couldn't be more exciting: from data we will get the cure for cancer as well as better hospitals; schools that adapt to children's needs making them happier and smarter; better policing and safer homes; and of course jobs. Data allows us to adapt and improve public services and businesses and enhance our whole way of life, bringing economic growth, wide-ranging social benefits and improvements in how government works.

This next phase of the digital revolution has PSI at the very foundation. Therefore Britain enjoys significant advantages: the size and coherence of our public sector (who else has critically important data of the range and depth of the NHS?) combined with government's strong commitment to a visionary open data policy means that we have the opportunity to be world leaders in the enlightened use of data. If we play it right we can break free of the shackles of a low-growth economy in which government and the public sector are seen as a resource drag and an obstacle, and they instead become key drivers of a transforming process.

Why is PSI so important? Consider the role of government: it exists to decide the rules by which people can act, and to administer them: how much, by what method, and from whom to take resources; and how to re-allocate them. Doing it well enables national success; doing it badly means national failure. Ensuring that the process of government is optimised for progress, and does not corrupt into an obstacle to progress, requires continuous data and the continuous analysis of data. To paraphrase the great retailer Sir Terry Leahy, to run an enterprise without data is like driving by night with no headlights. And yet that is what government often does. It has a strong institutional tendency to proceed by hunch, or prejudice, or by the easy option. So the new world of data is good for government, good for business, and above all good for citizens. Imagine if we could combine all the data we produce on education and health, tax and spending, work and productivity, and use that to enhance the myriad decisions which define our future; well, we can, right now. And Britain can be first to make it happen for real.

To do that we need to move faster and with even greater commitment to creating the essential infrastructure. We should realise that there is a difference between a commitment to transparency and a true National Data Strategy for economic growth. How we self-consciously develop and implement that strategy is the theme of this Review.

It is now time to build on the very positive start we have made on open data with a more directed, more predictable engineering of usable information. Obstacles must be cleared, structures defined, and progress audited, so that we have a purposeful, progressive strategy that we can trust to deliver the full benefits to the nation.

My recommendations fall into five basic themes:

- defining the principles of ownership: it all belongs to the citizen, not to the government
- creating a national data strategy for maximising our opportunity - a plan that is recognisable outside government, actionable, and auditable
- accelerating implementation so that delivery is broader and more reliable, and that data is utilized in commerce and public administration
- strategic focusing of support for the new infrastructure (including strategic investment in basic data science)
- ensuring trust in the system: confidentiality must be strengthened by fully deploying the available technology of data security, and imposing higher penalties for infractions

This review does not call for any significant increase in spending on a national data strategy, nor any additional administrative complexity; rather, it calls for a broadening of objectives together with a sharpening of planning and controls. We should remain firm in the principle that publicly-funded data belongs to the public; recognise that we cannot always predict where the greatest value lies but know there are huge opportunities across the whole spectrum of PSI; appreciate that value is in discovery (understanding what works), better management (tracking effectiveness of public administration), and commercialisation (making data practically useful to citizens and clients); create faster and more predictable routes to access; and be bold in making it happen.

From a Transparency Policy to a Growth Strategy

The digital revolution has already fundamentally changed how we live and work together. The creation of the Internet was about new platforms for communication and organisation, which allowed us to connect in new ways - to share information better and faster, to buy and sell things at greater distances and lower cost. More change, just as big, is coming as a result of further exponential growth in computing power. The world will look even more different twenty years from now than 1993 looks to us today.

The next phase is about using new information to change how we make decisions. You can already see it in the way we travel: live information about every detail of our transport systems means we don't have to guess when the next bus will arrive or the most efficient route from A to B, a development that has been estimated to have generated a value of £15-58 million each year in saved time for users of Transport for London. The next big leaps forward, both in improving our lives and creating national prosperity, will be in data-driven medicine, education, more effective allocation of resources, and economic development.

Underway today is a huge increase in the amount of structured data which we are producing through our everyday activity and, crucially, our capacity for storing it and crunching it (that is, using computers to turn data into usable information) and making it part of our daily processes of living and working. So we are not only seeing an explosion of data but an acceleration of how it can be converted into life-changing tools: the constant natural production of data combined with advances in science and engineering, especially in machine learning, artificial intelligence and robotics, means that we will be able to have self-driven cars, lower-cost lower-error operations, life-enhancing health regimes, a great expansion in learning, and the targeting of investment by continuous evidence of outcome to get the greatest benefit to all citizens without the obstacle of political guess-work.

It's a grand claim, which can best be validated outside the scope of this review. The accompanying claim that Britain can be the world leader in this revolution can easily be made credible here by reminding us of three self-evident conditions: 1) Britain has had well-developed administration systems, delivering significant welfare support and public services, since the second world war; 2) Britain is in the front rank of scientific and engineering excellence, with the highest quality of universities, our knowledge base is the most productive in the G8; 3) Britain is already at the forefront of the open data movement, being ranked 1st on the European PSI scoreboard, with a policy that has been driven and supported by all the major political parties.

As an example of innovation, we have the Open Data Institute.[1] Co-funded by government and business, the ODI is well-placed to demonstrate the value latent in PSI, for example through building the demand side for PSI (including public sector use of its own data and incubating start-ups), and training business to best exploit and innovate with the data released by government. It will be one of the key contributions to developing our capability. We will also have to look to how we focus resources within academia. The massive increase in the volume of data generated, its varied structure and the high rate at which it flows have led to a new branch of science being developed – data science. Many existing businesses will have to dramatically engage with "big data" to survive, but unless we improve our base of high-level skills few will have the capacity to innovate to create new approaches and methodologies that are simply orders of magnitude better than what went before. We should invest in developing real-time, scalable, machine-learning algorithms for the analysis of large data sets, to provide users with the information to understand their behaviour and make informed decisions.

So, the next phase of economic, scientific and social development has data as its core - the digital trace left by human activity that can be readily gathered, stored, combined and processed into usable material. This data, to optimise its value to society, must be open, shareable and, where practical, it should be free. The richest source of data is government, which accounts for the largest proportion of organised human activity (think health, education, transport, taxation, welfare, etc). Therefore Britain must focus intellectual attention and material resources on the task of fulfilling the potential of PSI. The benefits will be many including: transparency, accountability, improved efficiency, increased data quality, creation of social value, increased participation, increased economic value, improved communication, open innovation, and data linkage. Just imagine this applied to health, an area in which we are making significant advances. There is a significant amount of work ahead. For instance, at the moment health data

[1] http://www.theodi.org/

comes through a variety of unconnected channels and into many different silos. It is hard for researchers to gain access to its full value. Advances in technology not only now allow us to collect data at source in real time, but also enable more practical linkage and accessibility. Establishing ways to effectively link data should become a priority, with special attention being paid to how medical practitioners can both access data themselves, and also contribute the data they have collected.

We already have the strong foundations of an open data policy, above all in the work of the Transparency Board.[2] In this area, government has been activist, intelligent and committed, working with enthusiastic committees on development and implementation, but it is still some distance from being a true bankable plan for building an infrastructure sufficient to the scale of the opportunity. For example, the Transparency Board defined the Public Data Principles, the key one being Principle 13: "Public bodies should maintain and publish inventories of their data holdings".[3] But this has only happened in small pockets and it certainly is not routinely published as open data; we do not know what the national data stock is. This demonstrates the need for a more purposeful and progressive approach.

Contributing to this Review is the argument that even the shorter-term economic advantages of open data clearly outweigh the potential costs, an argument substantiated in many cases by the accompanying document prepared for this Review by Deloitte. Deloitte analysis quantifies the direct value of PSI at around £1.8bn with wider social and economic benefits taking that up to around £6.8bn. These are compelling estimates and undoubtedly conservative (I asked Deloitte to focus on the clearly-defined economic value, and avoid speculation about the undoubted acceleration of benefit with the new technologies coming on stream).

We build on this with an argument that we have already taken significant positive steps in releasing PSI and have learned from the process - and seen concrete benefits (some of which are described in the following pages) which should make us confident in considerably accelerating the process.

PSI is incredibly diverse and its lack of homogeneity presents challenges in explaining what we are referring to. For this review I have used the definitions of PSI enshrined in legislation. This is helpful in determining what is and is not included within scope. It is less helpful when we seek to apply universal principles which cover all uses of PSI.

> "PSI covers the wide range of information that public sector bodies collect, produce, reproduce and disseminate in many areas of activity while accomplishing their public tasks."

Source: adapted from BIS and APPSI Glossary

There have been reviews before this one which have made recommendations on sub-sets of PSI (such as research data, administrative data, health data) where, for very good reasons, 'open' cannot be applied in its widest context. I therefore suggest we acknowledge a spectrum of uses and degrees of openness. For example, with health

[2] http://blog.okfn.org/2011/10/21/transparency-board-urges-widest-possible-response-to-uk-data-consultations/

[3] http://data.gov.uk/blog/public-data-statement-of-principles

data, access even to pseudonymous case level data should only be to approved legitimate parties whose use can be tracked and against whom penalties for misuse can be applied; and access can be limited to the secure sandbox technologies - initiatives that give access to data to researchers in a controlled way, while respecting the privacy of individuals and the confidential nature of data. An example is the Economic and Social Research Council's (ESRC) Secure Data Service, providing access to de-identified research data. Under these conditions, we can greatly extend access to connected data that spans the whole health system, and to many more practitioners, and much faster, than is currently the case, with the result that we gain the benefits of 'open' but without a significant increase of risk. Nor should we consider 'free' ('at marginal cost') to be the only condition which maximises the value of PSI; there may be some particular cases when greater benefits accrue to the public with an appropriate charge.

This Review is based on what we have tried to make an exemplary process of true consultation. I have met with a wide variety of interested and informed parties through breakfast seminars themed around Big Data, Linked Data and Health, and other larger gatherings including with big businesses, SMEs and start-ups. I have also interviewed experts, activists and practitioners. Most important perhaps, I have run two waves of surveys, each with simple, defined multi-option questions, and with every question accompanied by an open comment box. The first wave was exploratory, helping us to develop our ideas, the second wave was confirmatory, seeking support for broad versions of the recommendations below. These surveys were run in two versions: an open format (in which anyone could take part online, with the link promoted across government and private enterprise communities via our email contact lists and to a broader audience from the YouGov Twitter account); and a closed format to a sample of the general population from the YouGov panel. Happily, all recommendations received overwhelming approval by both groups, and the final recommendations reflect suggestions from supporters as well as the few opponents of the recommendations. The full data from the two waves of surveys to both groups as well as the notes to the seminars have all been made openly available.

My recommendations do not (and were never intended) in themselves to constitute a plan – but, one might say, they outline a strategic approach, which should be rapidly be taken forward by government. Simply put, the strategy is:

A. Recognise in all we do that PSI, and the raw data that creates it, was derived from citizens, by their own authority, was paid for by them, and is therefore owned by them. It is not owned by employees of the government. All questions of what to do with it should be dealt with by the principle of getting the greatest value back to citizens, with input not just from experts but also citizens and markets. This should be obvious, but the fact that it needs to be constantly reaffirmed is illustrated by the way that even today, access to academic research that has been paid for by the public is deliberately denied to the public, and to many researchers, by commercial publishers, aided by university lethargy, and government reluctance to apply penalties; thereby obstructing scientific progress.

B. Have a clear, visible, auditable plan for publishing data as quickly as possible, defined both by bottom-up market demand and by top-down strategic thinking, overcoming institutional and technical obstacles with a twin-track process which combines speed to market with improvement of quality: 1) a 'early even if imperfect' track that is very broad and very aggressively driven, and 2) a 'National Core Reference Data' high-quality track which begins immediately but narrowly; and then moving things from Track 1 to Track 2 as quickly as we can do reliably and to a high standard. ‘Quickly’ should be set out by government through publicly committed target dates.

C. Drive the implementation of the plan through a single channel more clearly-defined than the current multiplicity of boards, committees and organisations that are distributed both within and beyond departments and wider public sector bodies. It should be highly visible and accessible to influence from the data-community through open feedback mechanisms. 'Implementation' includes not only publishing but also processes to ensure that government transparently uses its own structured data to improve policy development and to measure progress.

D. Invest in building capability for this new infrastructure. It is not enough to gather and publish data; it must be made useful. We lack data-scientists both within and outside of government, and not enough is being done in our education system at school and undergraduate level to foster statistical competence; we will feel these gaps more and more as the potential grows. Government is already committing resources to this; we should consider increasing this further, as the economic and social benefits quickly and demonstrably outstrip costs. Our research councils should seek to play a more strategic role, targeting investment on basic data-science and on inter-disciplinary academic/business projects and partnerships.

E. Ensure public trust in the confidentiality of individual case data without slowing the pace of maximising its economic and social value. Privacy is of the utmost importance, and so is citizen benefit. People must be able to feel confident about two things simultaneously: that the data they have supplied or that has been collected about them is made as useful as possible to themselves and the community; and that it will not be misused to their detriment. We lay out ways in which we think we can get as close as possible to this ideal.

Stephan Shakespeare

Summary of Recommendations

Recommendation 1

The government should produce and take forward a clear, predictable, accountable 'National Data Strategy' which encompasses PSI in its entirety. A significant part of the strategy should include the actions outlined in the Open Data White Paper[4], but it should also bring together other policy developments including the Finch Report[5], the Administrative Data Taskforce, the forthcoming Information Economy Strategy[6], and the Midata initiative[7], as well as the whole spectrum of PSI. The strategy should explicitly embrace the idea that all PSI is derived from and paid for by the citizen and should therefore be considered as being owned by the citizen. It is the therefore the duty of government to make PSI as open as possible to create the maximum value to the nation.

We already have strong beginnings of a PSI approach and enthusiastic committees for implementing it, but it is some way from being a true plan for building a governance and technology infrastructure sufficient to the scale of the opportunity. In our consultations, business has made clear that it is unwilling to invest in this field until there is more predictability in terms of supply of data. Therefore without greater clarity and commitment from government, we will fail to realise the growth opportunities from PSI.

It is important to note for such a strategy that the biggest prize is freeing the value of health, education, economic and public administrative data.

Detail: Government should work together with other parts of the public sector to produce a National Data Strategy that brings together existing policy and guidance. The national strategy should be defined top-down but build on engagement with data communities, implemented by a non-government departmental team, and audited externally.

Recommendation 2

A National Data Strategy for publishing PSI should include a twin-track policy for data-release, which recognises that the perfect should not be the enemy of the good: a simultaneous 'publish early even if imperfect' imperative AND a commitment to a 'high quality core'. This twin-track policy will maximise the benefit within practical constraints. It will reduce the excuses for poor or slow delivery; it says 'get it all out and then improve'.

The intention is that as much as possible is published to a high quality standard, with departments and wider public sector bodies taking pride in moving their data from track 1 to track 2.

The high-quality core should be enshrined as National Core Reference Data. It should

[4] https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential
[5] http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf
[6] https://www.gov.uk/government/speeches/industrial-strategy-cable-outlines-vision-for-future-of-british-industry
[7] https://www.gov.uk/government/publications/better-choices-better-deals-report-on-progress-on-the-consumer-empowerment-strategy

be defined top-down, strategically, from both a transparency and economic value point of view (and not, as now, by the departments and wider public sector bodies themselves). Within such National Core Reference Data we would also expect to find the connective tissue of place and location, the administrative building blocks of registered legal entities, the details of land and property ownership.

Appropriate metadata should wherever possible be published alongside data, so users know what the quality limitations are and therefore how and for what purposes it is appropriate to use the data.

Detail:

i) We should define 'National Core Reference Data' as the most important data held by each government department and other publicly funded bodies; this should be identified by an external body; it should (a) identify and describe the key entities at the heart of a department's responsibilities and (b) form the foundation for a range of other datasets, both inside and outside government, by providing points of reference and interconnection.

ii) Every government department and other publicly funded bodies should make an immediate commitment to publish their Core Reference Data to an agreed timetable, to a high standard agreed to maximise linkability (as far as is possible within the constraints of not releasing personally identifiable data), ease of use and free access. They should also commit to maintaining that dataset and keeping it regularly updated. The scope should also be extended to include wider public sector funded bodies and agencies.

iii) Alongside this high-quality core data, departments and other public sector bodies should commit to publishing all their datasets (in anonymised form) as quickly as possible without using quality concerns as an obstacle - that is, if there is a clash between data quality and speed to publication, they should follow the 'publish early even if imperfect' principle because data scientists are well accustomed to getting value out of imperfect data. Currently many datasets are held back because it is felt they are not ready because they are not of sufficiently high quality, and that resources prevent their speedy improvement. But data users say that lower quality is not as much of a problem as is non-publishing.

iv) This will require measured and incremental improvement. Therefore, government should commit to reporting annually on the progress that has been made to meet this twin-track policy. There should be a co-ordinated programme of audit for each department and public sector funded body of their open data performance with recommendations for further release. The system of departmental information asset registers should be standardised to make searching and navigation easier and should be expanded to include routine consideration of the suitability for publication of both structured and unstructured information.

Recommendation 3

There should be clear leadership for driving the implementation of the National Data Strategy throughout the public sector. There are many committees, boards, overseers and champions of data; but no easily understood, easily accessed, influential mechanism for making things happen. There should be a single body with a single public interface for driving increased access to PSI.

Supporting the leadership should be a “data intelligence and innovation group” to provide external challenge and aid delivery. This group, which may be linked to the ODI, should perform a non-executive role.

Detail: A review of current governance structures for PSI is needed to identify a primary channel to lead the implementation of the National Data Strategy, and the controls it can use to be most effective. This should be a simplification process, not an increase of governance complexity and it should increase the connectivity between boards/groups to limit duplication of effort and actions that are not aligned appropriately.

Recommendation 4

One would be hard-pressed to find any expert who, asked to create new structures for core reference data from scratch, would advocate the current Trading Fund model (for Companies House, Land Registry, the Met Office and Ordnance Survey) in today’s world of open data. One would question the current quasi-commercial Trading Fund model, in favour of one which would be responsible for high quality and transparent data production (that is, collecting and publishing data that is required by parts of the public sector to execute the public task, in a way that can be seen to be reliable and authoritative), publishing this as open data and engaging in activities beyond this only where they are confident that they will not crowd out private and third sector activity and innovation.

But we are not starting afresh, and we have, in the Trading Funds, organisations of high quality which one should hesitate to disrupt. The Met Office, for example, is a world-leading forecaster, a pioneering scientific institution that is already publishing vast amounts of data. It would be risky to stop it doing what it is good at and leaving it to others in the market to fill the gap - there would be clear risks to national resilience, including to lives and property.

That does not mean we should not press hard for significant adaptation of the model to the new potential for open data.

Each of the Trading Funds has an essential role in the collection, processing and maintenance of high quality core-reference data to enable the public sector to do its job and for maximum economic benefit. However, the current Trading Fund model is now out of step with the government’s open data aspirations.

Some good progress has been made in opening up data for public sector sharing and re-use. But restrictive licensing, applied to key PSI, limits the opportunity for businesses, especially SMEs, to make effective use of PSI as an underpinning business resource.

Detail:

i) The overarching aim of the Trading Funds should be to deliver maximum economic value from public data assets they provide and support, by working to open up the markets their data serves. This means they should work towards opening up all raw data components, under the Open Government Licence (OGL) for use and re-use.

ii) They should reconsider their product and service development activities in the light of a new era where they can potentially deliver greater economic benefit through improved joint-working with third parties.

iii) They should better communicate what data is available for use/re-use and how it can be used/re-used under the simplified licensing terms; building on their existing efforts to raise greater awareness amongst the user community.

iv) They should deliver more support for third-party users including the greater use of 'hack days' and data-user competitions to demonstrate the value of particular PSI datasets.

v) They should enable greater provision of 'sandbox' or secure online environments to allow users to explore datasets without prohibitive costs of entry or participation.

To promote and support a more beneficial economic model for Trading Fund data government should review how the Trading Funds are recognised and rewarded for their activities to stimulate innovation and growth in the wider markets they serve

Recommendation 5

We should have a clear pragmatic policy on privacy and confidentiality that increases protections for citizens while also increasing the availability of data to external users. We can do this by using the developing 'sandbox' technologies, or 'safe havens' as they are referred to by the Administrative Data Taskforce[8] and the Data Sharing Review[9], that allow work on data without allowing it to be taken from a secure area. Along with appropriate anonymisation,putting in place guidelines for publication that more obviously pushes responsibility for (mis)use on the end (mis)user, and greatly strengthens application of punitive consequences, is critical. Especially sensitive datasets should be accessible only to those who can demonstrate sufficient expertise in the area and whose activity with the data is traceable. But that accreditation process should then be broad and simple, as the sandbox technology means we can trace activity and hold individuals responsible for misuse.

Data should never be (and currently is never) released with personal identifiers, and there are guidelines that should be followed to reduce the risk of deliberate attempts to identify data being successful. No method, including traditional non-digital information storage, is proof against determined wrong-doers. We do not require builders to only build houses that cannot be burgled. We do our best and impose consequences on the

[8] Administrative Data Taskforce http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative-initiatives/Administrative-Data-Taskforce.aspx
[9] Thomas / Walport Data Sharing Review http://www.ico.org.uk/upload/documents/pressreleases/2008/thomas_walport_statement.pdf

burglar not the builder. We currently have an unrealistic degree of expectation of any data controller to perfectly protect all our data - an attitude that inhibits innovation. Following 'best practice' guidelines should be enough, so long as we are willing to prosecute those who misuse personal data. Otherwise we will miss out on the enormous benefits of PSI. We should encourage continuing vigorous debate to achieve the right balance between the benefits and risks of open data (including whether citizens might in certain cases be enabled to opt out of open data). In considering further legislation we should institute increased penalties – not only loss of accreditation and much heavier fines, but also imprisonment in cases of deliberate and harmful misuses of data.

And we should be respectful of personal confidential data and follow the principles set out in the Information Governance review chaired by Dame Fiona Caldicott[10].

Detail:

i) Government should provide clear guidelines to all involved, whether data controllers, data holders or data users, that set out the approved ways of making data open and that if these guidelines are followed, liability for mis-use falls on the mis-user; also defining what constitutes a misuse of data or breach of privacy.

ii) The current complaints procedure for instances of data misuse should be made more accessible and awareness of the procedure should be improved.

iii) There should first be an assessment of existing guidance tools. Organisations should be encouraged to make greater use of Information Commissioners Office Codes, as a framework to develop their own policies, as well as using Privacy Impact Assessments (PIAs) as a flexible way to assess risks. Data.gov.uk should be updated to include an online guide of procedures and processes that apply to all public sector organisations, to improve clarity and awareness of information of help available and ensure that all organisations are working to the same guidelines. The guide should complement the Government Digital Service Service Design Manual, which includes information on procurement of data-release friendly IT, licensing, technical advice and standards.

Recommendation 6

Building on existing activities around capability, there should be a focused programme of investment to build skill-sets in basic data science through our academic institutions, covering both genuinely unfettered 'basic research' and research of 'practical immediate value' to the national data strategy. We cannot rely only on markets and government departments and wider public sector bodies to maximise the potential of this relatively new and fast-developing field in which we are positioned to be a world leader.

At the moment, the USA invests massively more than us and continuously reaps the benefits in world-leading business applications of science and technology; yet Britain is capable of being first in this field, given our expertise in data science and the fact we have large, coherent datasets. For example, nowhere in the world has such good health

[10] To share or not to share. https://www.gov.uk/government/publications/the-information-governance-review

data, due to the scale of the NHS as a single provider. There is huge potential here for building social and economic value if we are willing to invest smartly.

Detail:

i) Traditional training will of course continue to play an important role, as well as interactive and workshop sessions - such as mash-up days - especially those involving external developers. These are useful for sharing knowledge and expertise and creating an environment which is conducive to experimentation and innovative thinking.

ii) Public sector organisations should consider how they meet their current and future skills needs to deal with the increasing availability and use of data from across the public sector.

iii) Government should explore solutions that can be implemented quickly to improve the skills base to be able to effectively manipulate and extract value from PSI.

iv) In addition, government should promote and support building capability amongst graduates. Government should task the research councils to be strategic in their funding of graduate training to encourage the growth of basic data science and inter-disciplinary projects, and consider further increasing funding available for teaching of data discipline.

Recommendation 7

We should look at new ways to gather evidence of the economic and social value of opening up PSI and government data, and how it can be further developed taking into account the latest innovations in technology. This evidence should be used to underpin a bold strategy of investment in an infrastructure of data in order to make the UK the world leader in this field, thereby gaining the greatest advantage in this new wave of the digital revolution.

Currently we can measure the costs of producing and publishing data, but we have no model for evaluating the economic or social benefits 'downstream', and so we may be undervaluing these activities, leading to under-investment of resources.

Detail: We should create a “data intelligence and innovation group” that includes experts from within and outside government that as part of its wider role supports, challenges and takes forward thinking on how to improve the collection, processing and use of PSI. One of the initial tasks for the group should be to provide independent advice on the methodological challenges and evidence gaps identified by this review, and develop proposals to address them. A further task of the group should be to fully embed an analytics approach within policy making.

Recommendation 8

We should expect systematic and transparent use of administrative data and other types of PSI in the formulation, implementation, monitoring and adaptation of government policy and service delivery, and formally embed this in the democratic process. PSI should be as much a part of the evidence base as evaluations and survey data. This should include information derived by third parties in the delivery of services funded by the public sector.

Although Government does use and publish some PSI as part of programme evaluations and in impact assessments, practice varies, and the wider consultation process is not generally considered to be effective. We should deepen and broaden the role of PSI in policy making.

Detail: Each government department and wider public sector body should review whether the PSI that they currently hold is being used to maximum effect in developing, evaluating and adapting policy. It should explain what data it used to support any new policy and above all what data will be collected (and published) for continuous measure of its effectiveness.

Recommendation 9

We should develop a model of a 'mixed economy' of public data so that everyone can benefit from some forms of two-way sharing between the public and the commercial sectors.

Where there is a clear public interest in wide access to privately generated data, then there is a strong argument for transparency (for example in publishing all trials of new medicines). As the Royal Society's *Science as an Open Enterprise* report sets out this warrants careful consideration in each case so that legitimate boundaries of openness are respected. For example, data could be made public after intellectual property has been secured or after a particular product has been launched. Where the data relates to a particularly and immediate public safety issue, it should be published openly as soon as possible[11].

A company working with government should be willing to share information about activity in public-private partnerships, as information about activity in public-private partnerships held by private companies is not currently subject to the Freedom of Information Act. This could be greatly enhanced without the need for legislation by creating a field in procurement forms asking for the company's open data policy regarding the sought contract.

Data that is derived from the activity of citizens must be seen as being at least co-owned by them and returning value to them, though the investment of business in collecting and processing the data should also be respected. There are government initiatives such as Midata, a government led project that works with businesses to give consumers better

[11] Science as an Open Enterprise, Royal Society 2012 http://royalsociety.org/policy/projects/science-public-enterprise/report/

access to the electronic personal data that companies hold about them. The project recognises that data about citizens belongs to them and that they should have a way of claiming and using their ownership. Midata is currently about empowering consumers – government itself should explicitly embrace the Midata initiative to empower citizens by returning key data it holds on citizens back to them.

Detail: Each government department should develop opportunities and regularly review the potential for two-way sharing between the public and commercial sector in the policy areas for which they are responsible.

Introduction

The scope and questions addressed by my review are outlined in the published terms of reference.[12]

What is the aim of the Review?

The review considers the full breadth of the PSI market, both current and future. It deals with the private sector, civil society and general public use and re-use of public information as well as the potential benefits for how the public sector uses and re-uses its own data. The review covers the elements set out below and includes answers to the questions posed.

The review establishes and takes stock of the current use and re-use of PSI within Government, making recommendations for improvements where appropriate. It will consider the current and anticipated future needs for Government given the current policy objectives across departments and wider public sector bodies as well as the opportunities and challenges presented by rapidly developing technology in the area.

In addition to the stated terms of reference, I have identified a number of further strategic questions that I have considered in this review:

- What types of PSI offer the greatest business opportunities?
- What are the biggest obstacles for government in order to unlock PSI opportunities?
- What might be done by Government to deal with any obstacles?

Who is the Review for?

My review is a call to Government to continue what has begun but at much greater pace and with increased focus. It represents my views which have been shaped and reinforced by those of many others across the country - from those already working with data to the citizens - who have all helped with the Review.

How was the Review carried out?

This has been a truly inclusive process. I have sought to ensure that this isn't just a review from data experts or those with a vested interest but it is truly representative. We followed a traditional approach of looking at the evidence and commissioning Deloitte to carry out fresh analysis of the market. With that I had the basis to start to form my views. There have been breakfast seminars, larger events with big businesses, SMEs and start-ups. I have also interviewed individual experts, activists and practitioners. I have also been fortunate to draw on the experience of my Data

[12] Published draft terms of reference: https://www.gov.uk/data-strategy-board#the-shakespeare-review.

Strategy Board colleagues. But my own evidence has come from the two waves of surveys, each with simple, defined multi-option questions, with every question accompanied by an open comment box. The first wave was exploratory, helping to develop ideas; the second wave, confirmatory, seeking support for my broad recommendations. These surveys were run in two versions: an open format (in which anyone could take part online, with the link promoted across government and private enterprise communities via our email contact lists and to a broader audience from the YouGov Twitter account); and a closed format to a sample of the general population from the YouGov panel.

How is the Review organised?

The Review provides an overview of the evidence base provided in the Deloitte Market Assessment of PSI, as well as the findings from the two surveys run by YouGov as part of the consultation process. Further chapters are structured according to the three key themes I have identified as priority areas: ownership, privacy and capability. The Review also contains detailed case studies on the current landscape of education and health data (Annex 1.2 and 1.3).

1. Evidence

In my foreword, I highlight the value of PSI in both economic and social terms. This chapter summarises some of the key evidence that is relevant to this review. The Market Assessment Report by Deloitte provides a wide-ranging analysis of the market for PSI. This includes an examination of the size, reach, and nature of the market for PSI. YouGov undertook a survey of public opinions. Its findings help us to understand what information people are interested in, how they use it, and their policy preferences. The chapter closes with a series of recommendations.

What is the value of Public Sector Information?

As part of the Shakespeare Review, Deloitte was commissioned to produce a market assessment report. This provides up-to-date evidence on the nature of the PSI market, its size, the types of people and organisations involved, the market's competitiveness and how we can best measure its performance.

The economic analysis undertaken by Deloitte for this review contains quantitative estimates on a number of measures of value. Key findings from the Deloitte Market Assessment these figures suggest an overall impact of around £6.8bn a year.

This figure comprises direct economic benefits estimated at around £1.8bn, and a wider social value of PSI conservatively estimated in excess of £5bn.. These direct economic benefits can be broken down into:;

- Consumer surplus from direct use and consumption PSI related products of around £1.6bn per year

- Producer surplus from revenue to PSI holders from sales of data of around £100m

- Supply chain effects from increased jobs and related consumer spending from the production of PSI of around £100m

The figures above have been subject to a 'sensitivity analysis' to model some of the uncertainties involved in estimation. This provides our central estimate with a range of £6.2bn to £7.2bn.

The economy wide estimates of the 'Wider Social Value' are augmented with detailed case studies which provide illustrations of where and how value is generated. For example, time saved as a result of access to real time travel data from Transport for London is valued at £15-58m.

.

What do people think about Public Sector Information?

It was important for the review to find out what views people hold about PSI – whether users and re-users of PSI, individuals or institutions. Knowing what people think will help to make sure that government fulfils their needs, and only acts where it should and can make a difference.

Some studies of people's attitudes to PSI have been published before. The European Commission most recently did so in 2011[13] while developing its policies on PSI. In that consultation, nearly nine out of ten respondents from a mix of sectors thought that re-use of PSI had not yet reached its full potential in Europe. The overwhelming majority of respondents thought that further action facilitating PSI re-use could help to unlock innovation.

The Open Data Dialogue[14] report, sponsored by Research Councils UK and JISC, which was published in June 2012, explored public views on open data, data reuse and data management policies within research. The study involved holding structured group discussions with around forty members of the public. Key findings emerging from this qualitative study included:

- Support for open research data, where it could improve people's health or the environment, and so was clearly in the public interest.
- Caution about open research data if it created a potential for harm, such as by encouraging poor decision making.
- Caution about privacy and confidentiality implications – such as lack of clear ownership of linked datasets.
- Public awareness that the incentives of researchers, companies and other parties were different, so that there was a need for governance.
- Open data could promote trust in general, but might not cause people to trust any particular interpretation of data more than another.

The Administrative Data Taskforce Report[15] considered public attitudes on the interaction between openness and privacy when providing access to administrative data for research purposes. It suggested that most of the well developed evidence on attitudes came from the health sector. The report summarised these attitudes by saying, "There is broad, though not unconditional, support for uses of administrative data for research. However, several studies do suggest the importance of demonstrating the value of such uses of data."

In order to add to the existing evidence base on public attitudes and to addresses specific questions relevant to the review YouGov, undertook a survey of public opinion between 22 February and 15 March 2013 about PSI [16] which was published on 2 April 2013. The survey asked the same set of questions around use and re-use of PSI to two different groups of people: a sample of the general public derived from a nationally representative YouGov sample, generating 777 responses (a sub-sample of people interested in this area screened from a 4,000-strong representative sample); and an online Open

[13] http://ec.europa.eu/information_society/policy/psi/docs/consultations/cons2010/results_online_consultation_final.doc
[14] http://www.rcuk.ac.uk/documents/documents/TNSBMRBRCUKOpendatareport.pdf
[15] http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative-initiatives/Administrative-Data-Taskforce.aspx
[16] http://datahub.io/dataset/shakespeare-review

Response survey, for anyone who wanted to share their views and ideas on PSI, which attracted 635 participants.

Figure 1.1 Levels of Familiarity with Data Issues

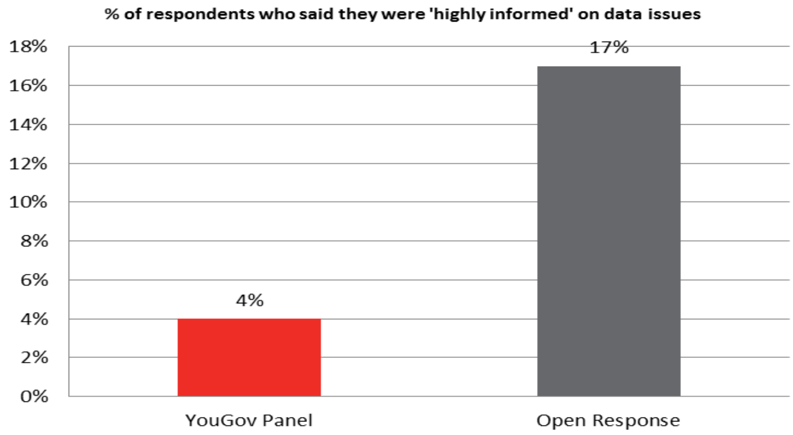
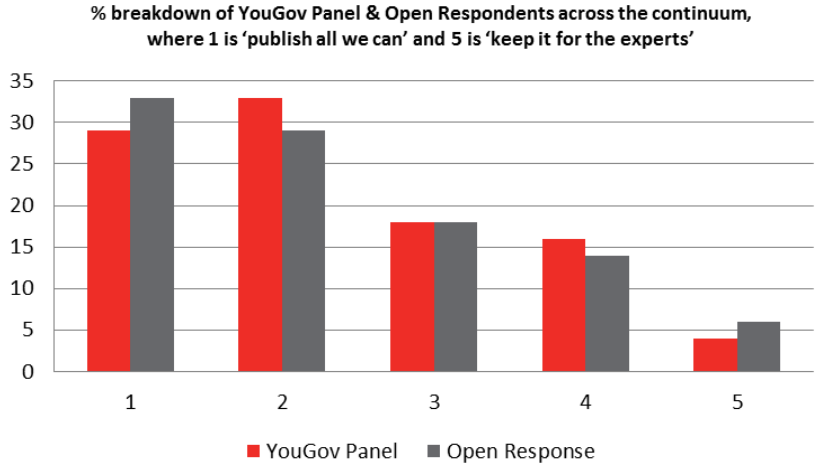


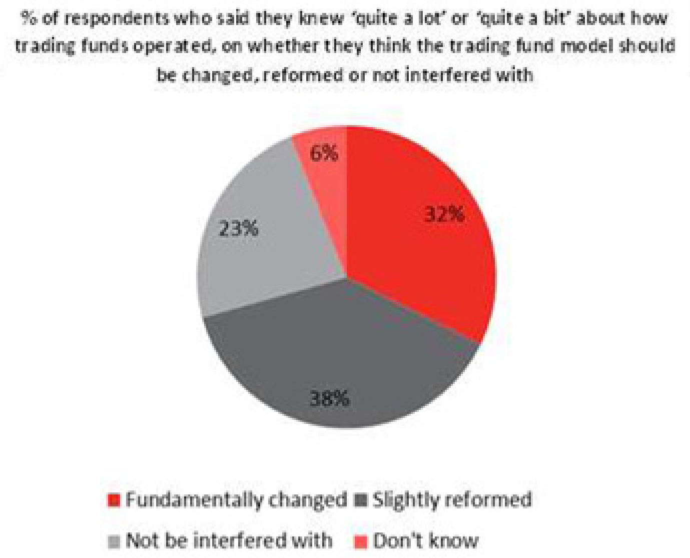
Figure 1.1 above shows that the two groups had different levels of familiarity with data issues. Of the YouGov Panel, 21% of respondents held a professional, academic or other special interest in open data. For the Open Response this figure was 46%.

Figure 1.2 Respondents' appetite for data publication



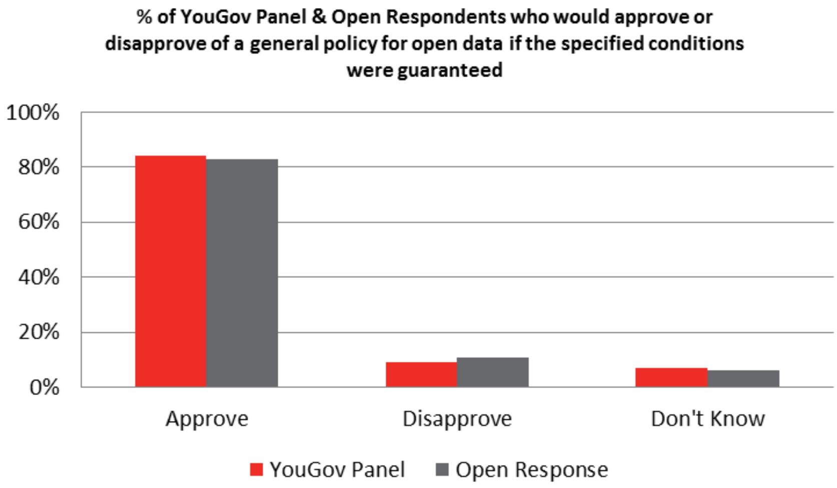
However, YouGov analysis suggests that despite this difference in backgrounds, the two groups held consistent opinions on many questions of content. For example, the two groups had similar levels of interest in information topics across health, education, police and geographic data and differing only on their interest in legal, government spending and communications data. The two groups held similar opinions on how much data should be published. 70% (68% YG, 74% OR) of total respondents think that we should make public all that we can about our health care system, while 25% (27% YG, 23% OR) believe we should just keep it to the experts to prevent confusion.

Figure 1.3 Respondents' Views on the Trading Fund Operating Model



On business models, respondents were asked 'how much they knew' about the Trading Fund model. YouGov asked those who said they know 'quite a lot' or 'quite a bit' (15% in total) whether they thought that the Trading Fund model should be changed, reformed or not interfered with at all. The opinions of the two groups of respondents are shown above.

Figure 1.4 Respondents' Views on a General Open Data Policy (with defined safeguards)



The survey also probed people's opinions on the interaction between opening up data and privacy. Respondents were asked whether they would approve or disapprove of a general policy for open data if certain conditions were guaranteed. The conditions were: anonymisation and pursuit of technical safeguards against personal identification; new rules on data misuse and stricter enforcement of existing rules; and that citizens could opt out of having their data published even under the conditions. In this scenario the two groups had almost identical opinions, with 83% (84% YG, 83% OR) of respondents stating they would approve of a general policy for open data and 10% disapproving (9% YG, 11% OR).

The study took the opportunity to ask people for their ideas on what is required to make the most of the potential opportunities PSI creates. YouGov analysis of the responses suggested that, “As you would expect a key theme to emerge was ‘education’. Other themes included utilising the skills and expertise of the private sector, as well as inspiring people by demonstrating the potential social and economic benefits.”

The research also asked people whether they thought the private sector should share more of its data with the public sector. The response indicated that around 24% of the YouGov panel and 17% of the Open Response thought that it wasn’t practical to expect commercial companies to share their data. On the other hand around 66% of the YouGov Panel and 77% of the Open Response thought that there were areas where companies should be made to share more data with the public.

To summarise, the key findings from the survey are as follows:

- The public may support opening up certain data, contingent on the safeguards of anonymisation, opt-outs and new penalties for misuse being put in place.
- PSI is an area of opportunities, but to make the most of them will require investment in skills, partnership with business and a focus on quality.
- Broad support for the idea of a reciprocal arrangement where companies were made to share more of their data with the public sector.

What are the key challenges and gaps in relation to the evidence base?

Rapid technological and recent policy changes pose new challenges to the evidence base for PSI. Smartphone apps, cloud computing, and growing internet connectivity have all contributed to the emergence of new distribution channels for PSI. Government policies on Open Data have led to greater availability of data and lower costs for re-users of it. Hence the Deloitte Market Assessment adds value to the evidence base by updating our knowledge of the market and how it works.

The Deloitte Market Assessment identified a number of evidence gaps, which if resolved could streamline analysis of PSI related issues. Three key gaps in availability and use of evidence are around accessibility, value and approach. These challenges are discussed below, together with potential solutions.

Accessibility

Records of which datasets government holds and publishes could be improved. This would help people find PSI and help government to evaluate its impact. One way to help achieve this would be to ensure that the system of departmental information asset registers includes routine consideration of the suitability for publication of both structured and unstructured information. This would help to embed a culture where the publication of information was thought about as a matter of routine.

Value

Better evidence is needed on who uses PSI and the value they attach to it. The importance of this need is underlined by the growth of open data – and the challenges that it poses to traditional analytical approaches which base value on prices. Specialist survey work offers a potential remedy here, and methods could be borrowed from other policy areas[17]. A panel of experts should be convened to advise on the best way forward for this vital work, since the results will be needed to inform a bold programme of investment in public sector information.

Approach

By fostering a greater climate of openness within government new opportunities opened up by technological change should be considered. Analytical data from the government's digital services could be fed back into policy making and service delivery. Insights into the formats and standards preferred by developers and which parts of society benefit most from public sector information could be derived from a variety of channels including analytics, user feedback and engagement with developers - for instance through the data.gov.uk forums or the Standards Hub which invites users to propose ideas about the data standards and formats that could best help to solve some of the government’s challenges around exchanging data. Metadata from government’s electronic services could be embedded within policy making. This could provide insights into the formats and standards preferred by developers and which parts of society benefit most from public sector information.

Recommendations

Recommendation 7

We should look at new ways to gather evidence of the economic and social value of opening up Public Sector Information and government data, and how it can be further developed taking into account the latest innovations in technology. This evidence should be used to underpin a bold strategy of investment in an infrastructure of data in order to make the UK the world leader in this field, thereby gaining the greatest advantage in this new wave of the digital revolution.

Currently we can measure the costs of producing and publishing data, but we have no model for evaluating the economic or social benefits 'downstream', and so we may be undervaluing these activities, leading to under-investment of resources.

Detail: We should create a “data intelligence and innovation group” that includes experts from within and outside government that as part of its wider role supports, challenges and takes forward thinking on how to improve the collection, processing and use of PSI. One of the initial tasks for the group should be to provide independent advice on the methodological challenges and evidence gaps identified by this review, and develop proposals to address them. A further task of the group should be to fully embed an analytics approach within policy making.

Recommendation 8

We should expect systematic and transparent use of administrative data and other types of PSI in the formulation, implementation, monitoring and adaptation of government policy and service delivery, and formally embed this in the democratic process. PSI should be as much a part of the evidence base as evaluations and survey data. This

[17] http://www.hm-treasury.gov.uk/d/green_book_valuationtechniques_250711.pdf

should include information derived by third parties in the delivery of services funded by the public sector.

Although Government does use and publish some PSI as part of programme evaluations and in impact assessments, practice varies, and the wider consultation process is not generally considered to be effective. We should deepen and broaden the role of PSI in policy making.

Detail: Each government department and wider public sector body should review whether the PSI that they currently hold is being used to maximum effect in developing, evaluating and adapting policy. It should explain what data it used to support any new policy and above all what data will be collected (and published) for continuous measure of its effectiveness.

2. Ownership

The diversity of PSI and the variety of organisations who collect, create and own data either as the public sector or on behalf of the public sector, is extensive. Ownership and the terms on which it is made available also vary. I think the time is now right to reflect on how the current models of ownership apply in the current context. The chapter closes with a series of recommendations.

What are the current models of ownership?

All government departments, their agencies and local government are creators and users of PSI. Most PSI is released without charge. However, there are some public bodies who collect or create data and provide a service and therefore charging is permitted, mainly to cover the costs of creating and maintaining that data. The Public Data Group Trading Funds – Companies House, Land Registry, the Met Office, and Ordnance Survey – are acknowledged for the volume and quality of their data and hence the potential value that can be derived from it. It is worth noting that the four Trading Funds already make increasing volumes of data available as open data[18] and there continue to be moves in this direction. There are other agencies who charge for their data[19] and therefore it is not sensible to treat all charging organisations as the same.

Similarly, if we are looking at where the real economic value may come in the future then I think there are far greater opportunities from other types of data, such as health and education. We have already seen great benefits to consumers and businesses from transport data which is now free and transforming our lives, making it easy to access all sorts of information quickly. This is an excellent example of where the market has stepped in to provide value-added services where little existed. This demonstrates a potential for other un-tapped markets.

Funding models

Charging and funding models for PSI are highly contentious and extremely complex. There remains an element of subjectivity as to what constitutes a dataset and what constitutes a 'value-add' service – with some data owners arguing that what is being charged for is not the PSI itself, rather its interpretation and analysis. Equally, as mentioned earlier, to treat all public sector organisations as the same is flawed.

Even if all PSI was made available free, the cost of generation, collection, retention and dissemination of the datasets must be funded. There are choices on where that funding comes from and this also introduces questions around pricing for access to public data.

It is worth bearing in mind that there are different costs across the different stages of the data lifecycle. As the Deloitte study sets out, for data collected as a by-product in an organisation's day-to-day activities, the marginal cost of its generation will be very low compared to the activity that caused the data to come about. However, the marginal cost

[18] For example, see www.ordnancesurvey.co.uk/oswebsite/products/os-opendata.html
[19] For example, the Office of National Statistics has a subscription charge for certain datasets.

of its dissemination to the wider public may be higher due to additional costs in formatting the data for use and to support re-users of the data.

As I mention earlier, the majority of data created from public funds should be freely available, however we must accept that charging can be legitimate. The key arguments are set out in the table below:

The case against charging	The case for charging
Charges for datasets create barriers to entry and expansion for SMEs and individuals to develop new products and services	**Aligning a revenue stream with a particular dataset will 'protect' it from any reductions in funding**, allowing data owners to continue to supply this even if they themselves must make other savings
The **charges prevent SMEs and individuals from 'experimenting' with the datasets** before they purchase to see if they are able to derive value from them, thereby making it hard to develop business cases	**A price can be interpreted as a signal of consumers' willingness to pay for a particular dataset's quality** and a commitment by the data owner to maintain this and offer support
Any **lost revenues to data owners from releasing datasets for no cost will be recovered by the Exchequer** in the long-run through increased tax revenues and more jobs being created	**Charging for certain datasets is necessary given they include elements of commercial or international datasets.** Where the benefit of a dataset being open is broadly spread, it is probably right for it to be free; when there are a narrowly limited set of beneficiaries, there is an argument for charging.

So I agree that some charging may be necessary. Where I differ from the current model is where the public sector is involved in value added services. The data revolution is moving rapidly, and faster than government structures are reacting to that change. I fully accept that when the Trading Funds were created that there was a demand for the wider products and services they delivered. However it is clear that the time is now right for the Trading Funds to step back and for the market to step up to create new growth by supplying value adding services.

The terms under which PSI can be charged are fairly clearly set out in Treasury guidance in managing public money but the direction of travel is firmly towards encouraging greater use of the Open Government Licence. This anomaly should be addressed.

The significance of Trading Funds

By far the greatest focus of opinion is on the four Trading Funds that make up the Public Data Group (PDG): Ordnance Survey, the Met Office, Land Registry and Companies House. The view has been expressed by some stakeholders that these data owners hold some of the most valuable datasets and there are strong arguments that these should be treated as core reference datasets available to all at no direct cost to the general public. Importantly, much of their data underpins the potential wider re-use and linking to other public or private datasets.

A high proportion of data from the four Trading Funds is already available as open data. However, despite these positive steps, there remains a perception among many consumers and commentators[20] that they are unable to access certain datasets for reasons of cost and this is creating a barrier to business growth. A number of studies[21] have argued that releasing these datasets as open data will have significant welfare benefits.

It is also worth bearing in mind that the PDG Trading Funds are not homogenous and so the change towards open data will affect each in different ways. The statutory register-based organisations, Companies House and Land Registry, are relatively straightforward in their purpose and funding model where a fee is charged for registration. Their commercial activities are fairly limited. The Met Office is a largely science driven organisation which collects an unimaginably large amount of weather and climate data which it then interprets and provides advice to the public, Government and to international colleagues as well as delivering tailored services to business. Ordnance Survey as the national mapping agency of Great Britain, creates and maintains the definitive and authoritative geospatial data for the Great Britain. Both the Met Office and Ordnance Survey have a higher level of commercial income and compete in their respective marketplaces. Moving towards a more open data model will affect each in a different way and over different timescales and this should be recognised in reform plans.

The price of change

The reforms that I have suggested should not result in an unjustifiably high cost to Government but putting a price on that is for Government to do. Carrying out a robust cost-benefit analysis of different funding model options for public data is complex as organisations differ greatly in their purpose and how they operate. Deloitte found that the available data on costs incurred from collection and dissemination by Trading Funds is not sufficiently detailed to determine what products were generating revenue and value. Forecasting future benefits is also hard to predict. How businesses and individuals might use datasets in the future to generate new products and services and by implication impact economic growth, is equally unknown. However, on the positive side, convincing the Treasury of the case for change must be justified using HMT Green Book[22] guidance; any benefits from a change in charging structures should include not just increased tax receipts but wider social benefits and costs in terms of organisational impacts.

Deloitte were able to estimate the cost on Exchequer revenue of continuing to collect and disseminate Trading Funds' PSI in its current form, without charging for it, is in the order of £395 million on an annual basis[23]. As government would no longer need to purchase the PSI itself, the direct loss to the Exchequer on an annual basis is in the order of £143 million. This figure may be lower still if there are efficiency savings to be made if fewer dedicated sales and marketing resources are required by Trading Funds. It seems a straightforward decision to invest £143m to make Trading Fund data widely available is a relatively small price to pay to leverage wider economic benefits far exceeding this by orders of magnitude.

[20] See for example www.freeourdata.org.uk/ and other references in Appendix 3.
[21] e.g. Pollock, 2011 and others – see the Deloitte Market Assessment, Appendix 3.
[22] Available at www.hm-treasury.gov.uk/data_greenbook_index.htm
[23] This figure comprises of Trading Funds' operating surpluses and the cost of data collection.

As in any market change, there will be advantages for some and potential threats to others. However, the dymamic nature of the data market surely encourages us to be proactive in determining what steps to take in seizing growth opportunities rather than reacting to change around us. Those currently deterred by charges would benefit from reforms and conversely, organisations who are at an advantage in using their own proprietary information for commercial advantage, might find their competitive advantage diluted if more PSI is released. But in dynamic markets this happens all the time and is a stimulus for innovation and so business should embrace the change.

My conclusion is that to quantify the costs and benefits precisely from outside Government is difficult due to the many complexities, however, I think there is sufficient evidence to support the theory that the benefits far outweigh the costs to releasing, firstly data from the Trading Funds and secondly, PSI across the public sector.

Looking ahead

Almost all of my review is focused on increasing the availability of PSI, but there are also opportunities from opening up private sector data. There have been real transformational benefits from initiatives such as Midata where consumers now have access to their own information collected on them by retailers and others. That is a huge step in really empowering consumers to take decisions based on data that they themselves have generated. I'm sure that Tesco didn't design their loyalty card scheme with open data in mind but this has been a truly groundbreaking step in access to private sector collected information. It also opened up discussions on who actually owns the data but I won't go into that further now.

I also see future opportunities from greater collaboration between the public and private sector in data sharing to transform how Government operates.

An excellent example of this is the delivery of the Government's smart meters policy. Led by the Department of Energy and Climate Change (DECC), this will see smart meters installed in 40-50m households across the UK by 2019. The potential is enormous from smart meter data, which will combine consumer and business information with energy company data and will directly inform government policy. Allowing customers and businesses to become more 'energy aware' will have positive impacts on behavioural aspects of energy usage and peak time demand side management. It has the potential to enable users to reduce energy spend at peak times, a particular strain on the energy supply network. Other benefits include reducing the impact on the environment by reducing emissions thus helping the UK meet international obligations. The data capacity needed to collect, analyse and transmit this big data set is considerable and so the proposal that a data control centre is established to collate smart meter data from consumers and ultimately businesses, recognises the need to effectively manage the data. This represents a developing government approach on handling data and opening up this data for wider use and analysis in the future will be important to stimulate growth opportunities.

I'm sure there are also other examples of government looking for new ways to use its own data and that of others to deliver better outcomes, deliver better policy and stimulate growth from open data.

Recommendations

Recommendation 4

One would be hard-pressed to find any expert who, asked to create new structures for core reference data from scratch, would advocate the current Trading Fund model (for Companies House, Land Registry, the Met Office and Ordnance Survey) in today's world of open data. One would question the current quasi-commercial Trading Fund model, in favour of one which would be responsible for high quality and transparent data production (that is, collecting and publishing data that is required by parts of the public sector to execute the public task, in a way that can be seen to be reliable and authoritative), publishing this as open data and engaging in activities beyond this only where they are confident that they will not crowd out private and third sector activity and innovation.

But we are not starting afresh, and we have, in the Trading Funds, organisations of high quality which one should hesitate to disrupt. The Met Office, for example, is a world-leading forecaster, a pioneering scientific institution that is already publishing vast amounts of data. It would be risky to stop it doing what it is good at and leaving it to others in the market to fill the gap - there would be clear risks to national resilience, including to lives and property.

That does not mean we should not press hard for significant adaptation of the model to the new potential for open data.

Each of the Trading Funds has an essential role in the collection, processing and maintenance of high quality core-reference data to enable the public sector to do its job and for maximum economic benefit. However, the current Trading Fund model is now out of step with the government's open data aspirations.

Some good progress has been made in opening up data for public sector sharing and re-use. But restrictive licensing, applied to key PSI, limits the opportunity for businesses, especially SMEs, to make effective use of PSI as an underpinning business resource.

Detail:

i) The overarching aim of the Trading Funds should be to deliver maximum economic value from public data assets they provide and support, by working to open up the markets their data serves. This means they should work towards opening up all raw data components, under the Open Government Licence (OGL) for use and re-use.

ii) They should reconsider their product and service development activities in the light of a new era where they can potentially deliver greater economic benefit through improved joint-working with third parties.

iii) They should better communicate what data is available for use/re-use and how it can be used/re-used under the simplified licensing terms; building on their existing efforts to raise greater awareness amongst the user community.

iv) They should deliver more support for third-party users including the greater use of 'hack days' and data-user competitions to demonstrate the value of particular PSI datasets.

v) They should enable greater provision of 'sandbox' or secure online environments to allow users to explore datasets without prohibitive costs of entry or participation.

To promote and support a more beneficial economic model for Trading Fund data government should review how the Trading Funds are recognised and rewarded for their activities to stimulate innovation and growth in the wider markets they serve

Recommendation 9

We should develop a model of a 'mixed economy' of public data so that everyone can benefit from some forms of two-way sharing between the public and the commercial sectors.

Where there is a clear public interest in wide access to privately generated data, then there is a strong argument for transparency (for example in publishing all trials of new medicines). As the Royal Society's *Science as an Open Enterprise* report sets out this warrants careful consideration in each case so that legitimate boundaries of openness are respected. For example, data could be made public after intellectual property has been secured or after a particular product has been launched. Where the data relates to a particularly and immediate public safety issue, it should be published openly as soon as possible[24].

A company working with government should be willing to share information about activity in public-private partnerships, as information about activity in public-private partnerships held by private companies is not currently subject to the Freedom of Information Act. This could be greatly enhanced without the need for legislation by creating a field in procurement forms asking for the company's open data policy regarding the sought contract.

Data that is derived from the activity of citizens must be seen as being at least co-owned by them and returning value to them, though the investment of business in collecting and processing the data should also be respected. There are government initiatives such as Midata, a government led project that works with businesses to give consumers better access to the electronic personal data that companies hold about them. The project recognises that data about citizens belongs to them and that they should have a way of claiming and using their ownership. Midata is currently about empowering consumers – government itself should explicitly embrace the Midata initiative to empower citizens by returning key data it holds on citizens back to them.

Detail: Each government department should develop opportunities and regularly review the potential for two-way sharing between the public and commercial sector in the policy areas for which they are responsible.

[24] Science as an Open Enterprise, Royal Society 2012 http://royalsociety.org/policy/projects/science-public-enterprise/report/

3. Privacy

In my Foreword I highlighted the importance of ensuring public trust in the confidentiality of individual case data without slowing the pace of maximising its economic and social value. Privacy is of the utmost importance, and so is citizen benefit. People must be able to feel confident about two things simultaneously: that the data they have supplied or that has been collected about them is made as useful as possible to themselves and the community; and that it will not be misused to their detriment. This chapter sets out the current privacy context in relation to PSI, and the ways in which I think we can get as close as possible to this ideal. The chapter closes with a recommendation.

Current context

The current **legislative landscape** around privacy includes the Data Protection Act and EU directives. The Data Protection Act 1998, which was enacted to bring UK law into line with the EU Data Protection Directive 1995, required Member States to protect citizens' rights and freedoms, including their right to privacy, with regard to the processing of personal data. As such it plays a significant role in the PSI landscape, particularly with regard to sharing and publishing data which may have personal privacy implications.

The Data Protection Act covers any data which concerns a living and identifiable individual. Anonymised data is not considered personal data and is therefore not covered by the Act. In this respect, the UK differs from several other EU Member States. Where anonymisation is reversible, however, the data does fall within the scope of the Act.

Privacy and data protection issues are embedded in European law in the form of the Data Protection Directive and the Re-use of PSI Directive. These directives are in the process of being revised ahead of adoption during 2013. In both directives there is an attempt to strike a balance between the protection of personal information on the one hand, and making information available for use and re-use. Under the Re-use Directive, both in its current and amended form, personal information is exempt from being re-used. This is reflected in licence models such as the Open Government Licence which do not extend to the re-use of personal information.

In recent years, government has undertaken work to better understand how to reconcile the desire for open government with the privacy of individual citizens. Notably, the review conducted by Kieron O'Hara for the Cabinet Office in 2011 which examined the issues for privacy that were raised by the Coalition Government's transparency programme.[25]

In the last two years, the **Information Commissioner's Office** (ICO) has issued two codes of practice relating to the re-use of PSI and data sharing:

[25] https://www.gov.uk/government/publications/independent-transparency-and-privacy-review

- ICO Code of Practice on Data Sharing (2011)[26]
- ICO of Practice of Anonymisation (2012)[27]

Both Codes of Practice aim to break down challenges of applying the Data Protection Act into practical guidance. The ICO has sought to develop a practical approach to the privacy issues arising from open data. The Code sets out a framework to enable better decision making about anonymisation, but it is not a 'how to' guide. It acknowledges the benefits open data can bring and how anonymisation can help achieve those, while protecting privacy. Anonymisation is becoming more challenging and the risks must be properly addressed.

In addition to the Anonymisation Code, the ICO also set up and funded the UK Anonymisation Network.[28] The aim of which is to create a practical forum for sharing practice related to anonymisation.

The ICO continues to work with the public sector to enable better, risk based understanding of data sharing, with more work planned to translate the messages in the Data Sharing Code to senior public sector officials in 2013.[29]

The **Administrative Data Taskforce** has recommended greater research access to administrative data and data linkage between departments. At the same time, the Taskforce has also recommended that an agreed set of ethical standards should be produced, drawing on well-established ethical guidelines and covering the research uses to which administrative data may and may not be put. The Administrative Data Taskforce also acknowledged the importance of public opinions and attitudes in relation to administrative data, and has recommended a strategy to engage with the public, including communicating the benefits of improved access to and linking between administrative data, and the measures being enacted to minimise risks of disclosure and to prevent inappropriate use of such data.

There are a number of examples **new sandbox technologies** - initiatives that give access to data to researchers in a controlled way, while respecting the privacy of individuals and the confidential nature of data. For example:

- The Administrative Data Taskforce describes a mechanism for linking datasets while preserving anonymity.[30]
- The Office for National Statistics has a secure access facility with remote access, known as the Virtual Microdata Laboratory.[31]

[26] http://www.ico.org.uk/for_organisations/data_protection/topic_guides/data_sharing See page 7 for definition of data sharing2012.

[27] http://www.ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation

[28] www.ukanon.net. The network is run by four partners – University of Manchester, Southampton plus the ONS and the ODI

[29] Data sharing was also closely examined in the 2008 Walport/Thomas report: http://www.connectingforhealth.nhs.uk/systemsandservices/infogov/links/datasharingreview.pdf

[30] http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf. For a further example of data made available to researchers, see http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf.

[31] http://www.ons.gov.uk/ons/about-ons/who-we-are/services/vml/index.html

- HMRC operate an on-site secure access facility for research access to some of their datasets. This is known as the HMRC Datalab.[32]

Fair Data, is an initiative run by the Market Research Society, encourages organisations to improve best practice in the collection, use and retention of customer and personal data.[33] It is an independently audited, voluntary accreditation scheme, where organisations sign up to the 10 Fair Data Principles, which cover collection and use of data, and the importance of protecting all respondents from harm.[34] As well as encouraging best practice within organisations, the scheme aims to build consumer trust by tackling what many see as a lack of control in how their data is used. The 'Fair Data mark' is a consumer facing mark which is a guarantee that an organisation meets the Fair Data principles. The scheme is starting to initiate negotiations with some serious current and potential entrants into the data market.

What are the challenges?

Our perception of privacy has changed dramatically in recent years. The internet and all forms of social media and mobile technology have produced a modern connected society that makes accessing and sharing information much easier. Over the last decade, sharing personal information has become common place for many. However, the use and misuse of technology has fuelled privacy concerns, and a declining sense of trust around use of personal data.[35] While many are happy to give their information to companies to take advantage of discounts or share personal information on social media sites, there is also an increasing sense that our information may be being used for purposes of which we are unaware.

Recent thinking has suggested that we need an alternative strategy to ensure privacy.[36] Viktor Mayer-Schönberger and Kenneth Cukier have envisioned an alternative framework of "privacy through accountability" for the big-data age, 'one focussed […] on holding data users accountable for what they do.'[37] This shift would allow the value of PSI to be unlocked by enabling greater flexibility around use of data in innovative ways, while ensuring accountability is pushed on to the data (mis)user.

I firmly believe that there are significant benefits to be gained from sharing and allowing the re-use of PSI although clearly safeguards and mechanisms are in place around information about named individuals. We now have the opportunity to reframe the use of personal data so it is proportionate for current expectations and flexibility of use.

Privacy issues can broadly be divided into issues related to the sharing of personal data (identifiable and related to individuals) in that form, which the Data Protection Act applies to in full, and the disclosure of data in anonymised forms which carries a number of risks and difficulties. The two key challenges that I believe government needs to be tackle are:

[32] http://www.hmrc.gov.uk/datalab/
[33] http://www.fairdata.org.uk/
[34] http://www.fairdata.org.uk/10-principles/
[35] Highlighted in the World Economic Forum's recent report 'Rethinking Personal Data', http://www3.weforum.org/docs/WEF_IT_RethinkingPersonalData_Report_2012.pdf
[36] See Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data* (John Murray, London, 2013) p. 156.
[37] *Big Data,* p. 175, 173.

1) the need for stronger measures for misuse of data, and

2) a lack of awareness of the rights and responsibilities as a data controller, holder or user.

Stakeholder consultations for this review have indicated that many feel that penalties for misuse of data are not severe enough.

There have been a number of calls to increase the severity of penalties for obtaining or disclosing personal data contrary to section 55 of the Data Protection Act.[38] Some steps have been taken to enhance the powers of the Information Commissioner in tackling serious or sustained breaches of the Data Protection Act. The ICO now has powers to issue penalties of up to £500,000 on a data controller in instances of serious breaches, aimed at incentivising good practice by both public and private bodies in their handling of personal data. In addition, the ICO has repeatedly called for custodial sentences to be available under section 55, and a recommendation was also made on this by Lord Justice Leveson. There is further work required to address this issue fully.

Stakeholder consultations have also indicated that the Data Protection Act is frequently, and often unjustifiably, cited as a blanket reason to not share or reuse data, and that a lack of awareness of privacy legislation is acting as a barrier to maximising the potential of PSI.

The Deloitte Market Assessment of PSI report suggests that while ‘the current legislative and regulatory environment around PSI is not acting as a barrier to generating value and market development’; the barrier is one of perception and awareness. It is a lack of understanding of the application and aim of the legislation that is preventing more release and sharing of information.

The Deloitte Market Assessment report highlights two specific challenges around attitudes towards data release which need to be addressed:

- regulations such as Data Protection are sometimes used as a shorthand justification for not sharing PSI within the public sector, with PSI holders not always able to translate their awareness of their rights and duties into scenarios where PSI is released or shared, causing a barrier; and
- when a policy decision is taken not to release PSI datasets to the general public, the reasons are often not well articulated or the conditions attached to access the data for purposes of re-use are overly restrictive.

Recommendations

Recommendation 5

We should have a clear pragmatic policy on privacy and confidentiality that increases protections for citizens while also increasing the availability of data to external users. We can do this by using the developing ‘sandbox’ technologies, or ‘safe havens’ as they are

[38] The section 55 issues are summed up in this recent Justice Committee report, which also backed the call for custodial sentences: http://www.publications.parliament.uk/pa/cm201213/cmselect/cmjust/962/96205.htm#a8

referred to by the Administrative Data Taskforce[39] and the Data Sharing Review[40], that allow work on data without allowing it to be taken from a secure area. Along with appropriate anonymisation,putting in place guidelines for publication that more obviously pushes responsibility for (mis)use on the end (mis)user, and greatly strengthens application of punitive consequences, is critical. Especially sensitive datasets should be accessible only to those who can demonstrate sufficient expertise in the area and whose activity with the data is traceable. But that accreditation process should then be broad and simple, as the sandbox technology means we can trace activity and hold individuals responsible for misuse.

Data should never be (and currently is never) released with personal identifiers, and there are guidelines that should be followed to reduce the risk of deliberate attempts to identify data being successful. No method, including traditional non-digital information storage, is proof against determined wrong-doers. We do not require builders to only build houses that cannot be burgled. We do our best and impose consequences on the burglar not the builder. We currently have an unrealistic degree of expectation of any data controller to perfectly protect all our data - an attitude that inhibits innovation. Following 'best practice' guidelines should be enough, so long as we are willing to prosecute those who misuse personal data. Otherwise we will miss out on the enormous benefits of PSI. We should encourage continuing vigorous debate to achieve the right balance between the benefits and risks of open data (including whether citizens might in certain cases be enabled to opt out of open data). In considering further legislation we should institute increased penalties – not only loss of accreditation and much heavier fines, but also imprisonment in cases of deliberate and harmful misuses of data.

And we should be respectful of personal confidential data and follow the principles set out in the Information Governance review chaired by Dame Fiona Caldicott[41].

Detail:

i) Government should provide clear guidelines to all involved, whether data controllers, data holders or data users, that set out the approved ways of making data open and that if these guidelines are followed, liability for mis-use falls on the mis-user; also defining what constitutes a misuse of data or breach of privacy.

ii) The current complaints procedure for instances of data misuse should be made more accessible and awareness of the procedure should be improved.

iii) There should first be an assessment of existing guidance tools. Organisations should be encouraged to make greater use of Information Commissioners Office Codes, as a framework to develop their own policies, as well as using Privacy Impact Assessments (PIAs) as a flexible way to assess risks. Data.gov.uk should be updated to include an online guide of procedures and processes that apply to all public sector organisations, to improve clarity and awareness of information of help available and ensure that all organisations are working to the same guidelines. The guide should complement the

[39] Administrative Data Taskforce http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative-initiatives/Administrative-Data-Taskforce.aspx
[40] Thomas / Walport Data Sharing Review http://www.ico.org.uk/upload/documents/pressreleases/2008/thomas_walport_statement.pdf
[41] To share or not to share. https://www.gov.uk/government/publications/the-information-governance-review

Government Digital Service Service Design Manual, which includes information on procurement of data-release friendly IT, licensing, technical advice and standards.

4. Capability

In my Foreword, I made it clear that strategic focussing of support for the new infrastructure (including strategic investment in basic data science) is needed. Capability is one the issues that I feel needs to be addressed as a priority. Insufficient capability is one of the most difficult challenges we face, and one we urgently need to address for the UK to be able to retain its position as a world leader in release and reuse of PSI. We lack sufficient data-scientists both within and outside of government. Data scientists are vital as they have a specific set of skills that enables them to analyse, interpret, curate and communicate. Our investment should include strong targeted resourcing of basic data-science which would include: training in writing computer code, a foundation in maths, statistics and probability as well help with the development of communicating skills to present data. This chapter looks at the steps that have already been taken to ensure we have the right skills, the barriers we have yet to overcome and makes recommendations about how government should go about addressing the capability shortage. The chapter closes with a recommendation.

Current context

At school age, all students should have a basic understanding of where data comes from and how it is used to solve problems. Findings from the PISA tests [42]of children aged 15 year olds puts the UK as being not statistically significantly different from the OECD average, and the TIMMS survey [43]has England and Northern Ireland above the international average, including in understanding data (Scotland and Wales did not participate). While the UK 's performance is comparable with most other OECD countries, we are being out-performed by East Asian countries. Recent research shows[44] that 24% of adults (8.1 million people) lack functional numeracy skills. Functional numeracy equates with skills below Entry Level 3, i.e. what would be expected of 9-11 year olds. In view of this evidence, there is still room for improvement and we need to continue to enthuse students about statistics and computing.

An ever-growing number of young people spend their free time playing computer games or online gaming; an interest that could be used in the classroom to improve data literacy and programming skills. There are some coding initiatives in schools, but no uniform requirement for schools to have such initiatives. The freedom to develop their own curricula has meant some schools have started teaching coding. A number of informal learning initiatives are teaching students how to code. *Apps for Good*[45] is a programme across 40 schools which helps young people solve a social problem by learning about tech development and entrepreneurship. *Code Clubs*[46] are after-school clubs run across

[42] http://www.oecd.org/pisa/
[43] http://timss.bc.edu/timss2011/international-results-mathematics.html
[44] http://www.bis.gov.uk/assets/biscore/further-education-skills/docs/0-9/11-1367-2011-skills-for-life-survey-findings.pdf
[45] http://www.appsforgood.org/
[46] http://www.codeclub.org.uk/

a number of primary schools to teach children how to code. A further example is the *girlswhocode* initiative[47] which is popular in the USA and simultaneously addresses the issues of gender bias and skills shortage.

In terms of increasing capability there have been a number of recent developments. The Open Data Institute has been established and will demonstrate the economic, social and environmental impact of open data by working closely with the public and private sectors as well as academia to unlock both publication and consumption. The Open Data Institute carries out a number of activities, including training open data technologists and entrepreneurs. Work is under way on the development of an Open Data Postgraduate Certificate[48], anticipating the first cohort to commence in early 2014.

There are a number of initiatives that are encouraging the development of statistical skills which include:

- The Royal Statistical Society, with the support of the Nuffield Foundation, has launched its GetStats [49] campaign, aimed at encouraging people to improve their statistical literacy.

- Nuffield/ESRC/HEFCE £15.5m 5 year programme[50] which addresses quantitative skills shortages in social science undergraduate teaching

While the broad ambition of the recommendations from the Administrative Data Task Force [51]are about easing access for research purposes to linked microdata it does reflect on capability. The Administrative Data Research Centres that the Task Force proposed would offer training as well as access to data for researchers.

In terms of the civil service, the Cabinet Office recently published Meeting the Challenge of Change - A capabilities plan for the Civil Service[52]. The civil service includes specialist roles that are directly related to PSI such as: economics, social research, statistics, information technology and operational research. Heads of Profession are accountable for building the organisational capabilities that are needed for their specialism and for helping people to build their individual capabilities. To support the Heads of Profession in this role, in September 2013 the Government will introduce a new Civil Service Professions Council. The council will be a co-ordinating body, bringing the professions together to work as a coherent force and maximise their overall contribution to capability building.

[47] http://www.girlswhocode.com/
[48] http://www.theodi.org/excellence/pg_certificate
[49] http://www.getstats.org.uk/
[50] http://www.nuffieldfoundation.org/quantitative-methods-programme
[51] http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative-initiatives/Administrative-Data-Taskforce.aspx
[52] http://engage.cabinetoffice.gov.uk/capabilities-plan/

What are the current issues and barriers to building capability?

As part of the Deloitte Market Assessment of PSI study undertaken within this review, the evidence reviewed suggests a number of access barriers to maximising the value of PSI which includes a lack of skills and understanding to fully exploit PSI. A number of studies have recently been published contending that advanced economies face a skills gaps in so-called 'data scientists' which will impact on all sectors of the economy. Although statisticians and experts on quantitative analysis form an established part of the academic community, data scientists differ from these existing professions in a number of important ways. As well as being able to work with large volumes of structured and unstructured data, they are able to translate these analyses into policy and commercial-ready insights and effectively communicate them to a range of stakeholders, often using innovative tools and visualisations.

The massive increase in the volume of data generated, its varied structure and the high rate at which it flows have led to a new branch of science being developed – data science. Drawing on skills and knowledge from fields such as computer science, mathematics and statistics and business studies, data scientists model complex business and/or research problems. They bring a variety of skills to bear on the problems to be solved including; the skills needed to integrate and prepare large and varied datasets; advanced analytics and modelling skills to reveal and gain understanding of hidden relationships within data; knowledge of the context within which the problem is defined; and good communication skills to present results. Traditional forms of analysis (relational database tools, analysis of structured files, etc.) cannot cope with big data. New approaches are being developed (e.g. 'shared nothing' architecture, Hadoop* software solutions, server clusters and cloud storage) which require skills and knowledge that are evolving at a pace which currently outstrips the supply of people with the knowledge and experience required.

While there may be current gaps in the supply of 'data scientists', economic theory suggests that in the medium- to long-term, the number of data scientists is likely to increase, filling the supply gap and reducing the current wage premium. However, in the short-term, this may mean PSI is left under-exploited and associated value remains locked out. The general scarcity and increasing competition for these skilled workers can make it harder to construct the infrastructure for world class PSI and scale up efforts to exploit its value.

The evidence received as part of the Deloitte Market Assessment suggests that in the UK:

- there is increasing demand for individuals with a portfolio of skills able to manipulate quantitative data, present it in innovative ways and generate commercial and policy insights from it;
- many of the individuals performing these roles have no specialised training, but rather have learned on the job and / or have a science/computation/mathematics background;
- businesses rarely designate specific 'data scientist' roles; rather, such analyses are done across a combination of professions such as statisticians, economists, researchers, analysts, operational researchers , policy and commercial managers – a dedicated data scientist would embody elements of all these roles;

- SMEs, in particular, will tend to be more affected by the increasing digital analytics and business growth is likely to be held back as companies fail to optimise the use of open data.
- certain industries such as pharmaceuticals, financial services, professional services and retail are increasingly dependent on these skills sets and a shortage of them would reduce the UK's international competitive advantage.

In undertaking the review concerns were highlighted over a lack of skills and familiarity to work effectively with data. Public sector officials have a long history of using PSI to inform policymaking without having dedicated data scientists. However, Government does have dedicated people who collect, manage, curate, analyse, interpret and communicate data. The concerns that were raised related to cultural biases against using PSI from outside home PSIHs, as well as having the necessary skills to combine and manipulate Big Data and Linked Data. The skills gap and increasing competition for these skilled workers from the private sector makes it harder to construct the infrastructure for world class PSI. This shortage of data scientists also hinders efforts to scale-up PSI data analytics. There is a risk that in the short term that the public sector will be priced out of the market when competing for data scientists and this would have a significant impact on the use and re use of PSI in particular quality. .

While there is considerable evidence of gaps in the analytical disciplines, we should not forget the delivery side of making PSI available. The management and preservation of data related to library (information science) skills, is crucial and requires specialist skills on knowledge of the domain and skills to index data and manage the preservation of the data over the long-term. This role may be over-shadowed by the more eye-catching analytical skills but they are two-sides of the data coin.

Data science skills are acquired in various ways – through undergraduate and postgraduate courses in computing science, mathematics and statistics, sometimes coupled with experience in a business or research environment. While businesses and the higher education sector have recognised the need to develop what is termed 'data analytics' as a critical skill set, there is an underlying problem that is exacerbating a growing skill shortage in this field and which will hamper the ability of businesses and the research community to extract value and gain insights from big data. In terms of its international standing, the UK compares unfavorably with many other countries in terms of the uptake of these subjects within the school system. Additionally, the higher education sector has been slow in recognising the growing demand for data analytical skills, especially those needed to address the challenges posed by 'Big Data'. There are some initiatives underway which are reflected in this chapter but unless action is taken to tackle this problem, via further short term initiatives to encourage the development of relevant courses within higher education, together with long term plans to raise the status of data analytical skills throughout the secondary education of children and young adults, the UK will fail to reap the benefits from the mass of information produced within the digital society, losing any competitive edge it could gain.

Modern technology provides opportunities to acquire data on a scale and to a depth that would have seemed unimaginable even a few years ago. Access to this data is uneven and is much more than a question of having some databases. Tools that allow you to query and learn from this complex data are of critical importance. So too are the methods and techniques that allow for a systems wide interpretation of the meaning and relevance of emergent patterns – to discern the wider context so as to distinguish the signal from the noise. Commercial advantage and good government will require innovation. Looking at developments so far, it would seem clear that many existing businesses will have to dramatically engage with "big data" to survive, but that most likely much of this engagement will be swept away, as a few have the capacity to innovate to create new approaches and methodologies that are simply orders of magnitude better than what went before.

Popular examples include for example the emergence of Google, which when it arrived, was so much better than the alternative search tools that it swept them away. Most people can appreciate innovation in products: a new app that does something new and powerful that we can now value even though we did not have it before. But search engines had existed before. Innovation in how you handle and understand data, innovation in the methodology, can be of decisive competitive importance.

We should invest in developing real-time, scalable, machine-learning algorithms for the analysis of large data sets, to provide users with the information to understand their behaviour and make informed decisions. More broadly we need to invest in understanding the patterns and insights to be derived from the large and broad sweeps of data available to us via the Web and Internet.

Creating a context that encourages abstract technical innovation that actually engages with socially and corporately important data questions is a challenge that needs to be supported and managed. Mathematicians, statisticians, computer scientists, and engineers all have insights into big data and can make the difference between a challenge being totally impossible and straightforward. But they cannot act in isolation – to ask the right questions and to genuinely innovate they need engagement with real world problems and abstract thinking on a substantial scale. We should support more partnerships where there one partner has focused business style objectives working in engagement with a partner who is focused on curiosity driven research which explores in a much broader way a business critical area for innovation. When this works it can be very productive and allow significant benefits for both parties and so create a sustainable project.

Recommendations

Recommendation 6

Building on existing activities around capability, there should be a focused programme of investment to build skill-sets in basic data science through our academic institutions, covering both genuinely unfettered 'basic research' and research of 'practical immediate value' to the national data strategy. We cannot rely only on markets and government departments and wider public sector bodies to maximise the potential of this relatively new and fast-developing field in which we are positioned to be a world leader.

At the moment, the USA invests massively more than us and continuously reaps the benefits in world-leading business applications of science and technology; yet Britain is capable of being first in this field, given our expertise in data science and the fact we have large, coherent datasets. For example, nowhere in the world has such good health data, due to the scale of the NHS as a single provider. There is huge potential here for building social and economic value if we are willing to invest smartly.

Detail:

i) Traditional training will of course continue to play an important role, as well as interactive and workshop sessions - such as mash-up days - especially those involving external developers. These are useful for sharing knowledge and expertise and creating an environment which is conducive to experimentation and innovative thinking.

ii) Public sector organisations should consider how they meet their current and future skills needs to deal with the increasing availability and use of data from across the public sector.

iii) Government should explore solutions that can be implemented quickly to improve the skills base to be able to effectively manipulate and extract value from PSI.

iv) In addition, government should promote and support building capability amongst graduates. Government should task the research councils to be strategic in their funding of graduate training to encourage the growth of basic data science and inter-disciplinary projects, and consider further increasing funding available for teaching of data discipline

5. Conclusion

I set out in my Foreword that obstacles must be cleared, structures defined, and progress audited, so that we have a purposeful, progressive strategy that we can trust to deliver the full benefits to the nation. We already have the strong foundations of an open data policy. In this area, government has been activist, intelligent and committed, working with enthusiastic committees on development and implementation, but it is still some distance from being a true bankable plan for building an infrastructure sufficient to the scale of the opportunity. Once we have taken steps to address the barriers identified in my report we need to see the data strategy as a major opportunity for government to transform the way it works; my vision is for a government and wider public sector that is pioneering in its use of its own data to create growth and improve public services. The chapter closes with a series of recommendations.

Bringing it all together: A National Data Strategy

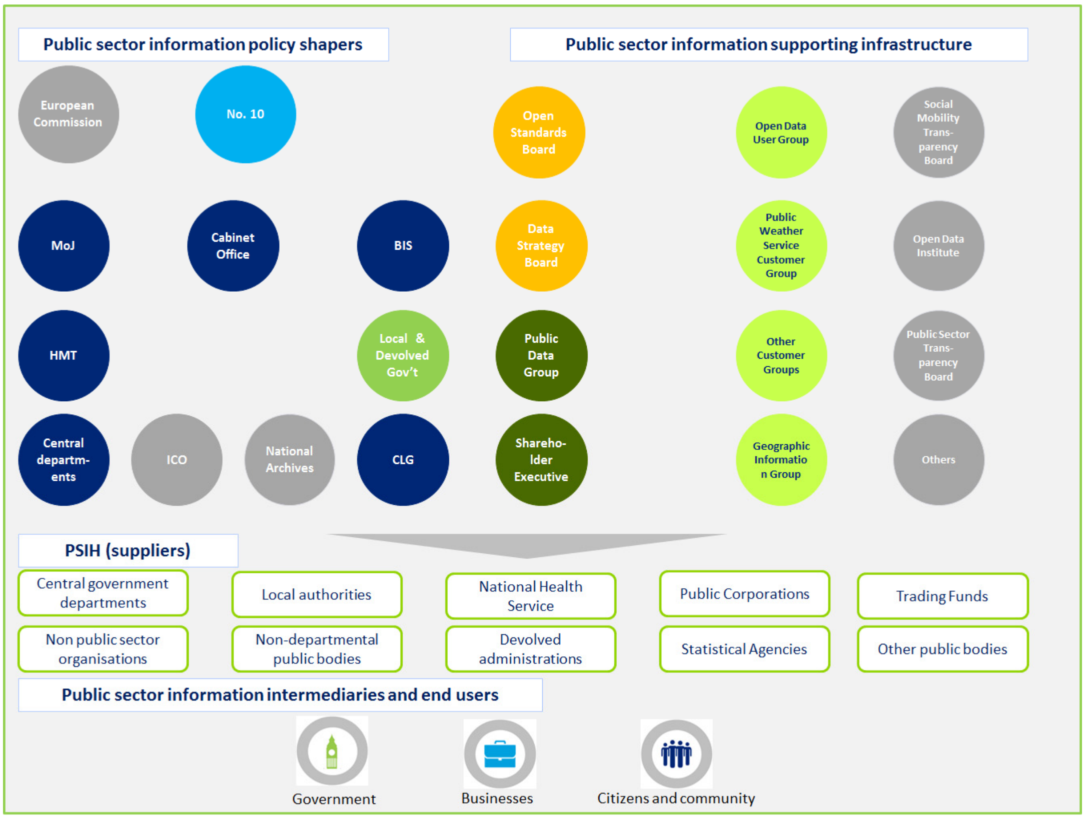
In my concluding remarks, I want to focus on the proposal in my report for a National Data Strategy and how we work together to develop and implement it. We have the constituent elements in place to build the National Data Strategy but we don't currently have a recognisable National Data Strategy. We need a more integrated approach, more clearly articulating the Government's PSI strategy. The National Data Strategy is about PSI but not just PSI. My recommendation is about reducing fragmentation and having both a strategic view of what we want to achieve as well a clear outline of what we are going to deliver when over the next three years. Progress needs to be made by then, otherwise we will seriously miss the growth opportunities.

How will we make it work?

It is encouraging to see that Government is engaging and investing in so much PSI-related activity. However, there is an obvious challenge. The number of bodies and groups involved presents a complex landscape to navigate, and there is a need to coordinate the disparate and on occasions overlapping activity across Government. The governance structure needs to be more easily understood by business and those outside central Government. The opportunity remains to streamline the PSI governance structures and processes to ensure that government has a clear and unified strategy.

As with any radical new policy approach, and particularly one as challenging as open data, the delivery structures tend to be designed to meet specific objectives and these evolve over time. In the case of the PSI market, the picture is complex with fuzzy boundaries. Deloitte presented an overview of the landscape which identifies the main players, but this is by no means definitive.

Figure 6.1 Overview of the Landscape



Source: Deloitte analysis. * Some policy shapers will also have operational duties, e.g. Cabinet Office is responsible for data.gov.uk

What this diagram doesn't show is the distinct responsibilities of each body/grouping nor the inter-relationships between them. To do so would make for an image that would be nearly impossible to fathom. It is completely understandable how this situation arose but there is a need for rationalisation to avoid the proliferation of further groups. In this world of shrinking resources, it is critical to ensure that effort is dedicated to delivering on the areas of greatest value and that duplication of effort is removed.

Streamlining is vital to stand any chance of success and the obvious place to start is with the clear leaders: Cabinet Office with its open data policy leadership role; Ministry of Justice with its lead on all legislation affecting PSI; the Transparency Board advising Francis Maude on transparency and open data; and the Data Strategy Board focused on identifying and brokering the release of public data. But with all the effort of these bodies, together with the myriad of other people involved, progress in delivering open data is still slower than expected.

So why is that? One observation is the nature of how Government works where departments have different roles and priorities. The incentives to deliver become diluted if the policy is not perceived to be central to their core objectives. I think that weakness is clearly present now. There is, of course, one unifying objective for Government which should sharpen focus and that has to be economic growth. I see a lot of smart thinking, good intentions and some instances of worthwhile delivery but, on the big ambition of open data, the follow through can be weak.

What would I do?

We need a shared view of the outcome we want to achieve whether we are a business, local authority or a government Department. We also need the infrastructure to deliver it, and some independent oversight to drive it. One way to achieve this would be to capitalise on the expertise of the Data Strategy Board which brings all these perspectives together to help drive the National Data Strategy. Working with the DSB and other experts, we could help build a strategy for the next three years. The strategy would be purposeful and progressive but also have specific commitments to what is going to be delivered in the first six months and beyond that. The Data Strategy Board or other similar group could act like the board of an innovative company, not playing an executive role but obsessed with maximising the opportunity of data for the nation, as if its share price depended on it and holding the rest of government to account. In streamlining the current governance structures, I would ensure there was a stronger focus on outcomes rather than single organisational responsibilities.

I want Britain to be like the West Coast of America in becoming the global focus creating a new world of social and economic growth driven by open data. All this would be mere arm-waving were it not for thee demonstrable underpinning facts: we have the expertise, we have the data, and we have cross-party, cross departmental, cross-sector consensus. Now let’s get the implementation system to make it happen.

Recommendations

Recommendation 1

The government should produce and take forward a clear, predictable, accountable ‘National Data Strategy’ which encompasses PSI in its entirety. A significant part of the strategy should include the actions outlined in the Open Data White Paper[53], but it should also bring together other policy developments including the Finch Report[54], the Administrative Data Taskforce, the forthcoming Information Economy Strategy[55], and the Midata initiative[56], as well as the whole spectrum of PSI. The strategy should explicitly embrace the idea that all PSI is derived from and paid for by the citizen and should therefore be considered as being owned by the citizen. It is the therefore the duty of government to make PSI as open as possible to create the maximum value to the nation.

We already have strong beginnings of a PSI approach and enthusiastic committees for implementing it, but it is some way from being a true plan for building a governance and technology infrastructure sufficient to the scale of the opportunity. In our consultations, business has made clear that it is unwilling to invest in this field until there is more predictability in terms of supply of data. Therefore without greater clarity and

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

[54] http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf

[55] https://www.gov.uk/government/speeches/industrial-strategy-cable-outlines-vision-for-future-of-british-industry

[56] https://www.gov.uk/government/publications/better-choices-better-deals-report-on-progress-on-the-consumer-empowerment-strategy

commitment from government, we will fail to realise the growth opportunities from PSI.

It is important to note for such a strategy that the biggest prize is freeing the value of health, education, economic and public administrative data.

Detail: Government should work together with other parts of the public sector to produce a National Data Strategy that brings together existing policy and guidance. The national strategy should be defined top-down but build on engagement with data communities, implemented by a non-government departmental team, and audited externally.

Recommendation 2

A National Data Strategy for publishing PSI should include a twin-track policy for data-release, which recognises that the perfect should not be the enemy of the good: a simultaneous 'publish early even if imperfect' imperative AND a commitment to a 'high quality core'. This twin-track policy will maximise the benefit within practical constraints. It will reduce the excuses for poor or slow delivery; it says 'get it all out and then improve'.

The intention is that as much as possible is published to a high quality standard, with departments and wider public sector bodies taking pride in moving their data from track 1 to track 2.

The high-quality core should be enshrined as National Core Reference Data. It should be defined top-down, strategically, from both a transparency and economic value point of view (and not, as now, by the departments and wider public sector bodies themselves). Within such National Core Reference Data we would also expect to find the connective tissue of place and location, the administrative building blocks of registered legal entities, the details of land and property ownership.

Appropriate metadata should wherever possible be published alongside data, so users know what the quality limitations are and therefore how and for what purposes it is appropriate to use the data.

Detail:

i) We should define 'National Core Reference Data' as the most important data held by each government department and other publicly funded bodies; this should be identified by an external body; it should (a) identify and describe the key entities at the heart of a department’s responsibilities and (b) form the foundation for a range of other datasets, both inside and outside government, by providing points of reference and interconnection.

ii) Every government department and other publicly funded bodies should make an immediate commitment to publish their Core Reference Data to an agreed timetable, to a high standard agreed to maximise linkability (as far as is possible within the constraints of not releasing personally identifiable data), ease of use and free access. They should also commit to maintaining that dataset and keeping it regularly updated. The scope should also be extended to include wider public sector funded bodies and agencies.

iii) Alongside this high-quality core data, departments and other public sector bodies should commit to publishing all their datasets (in anonymised form) as quickly as possible without using quality concerns as an obstacle - that is, if there is a clash between data quality and speed to publication, they should follow the 'publish early even if imperfect' principle because data scientists are well accustomed to getting value out of imperfect data. Currently many datasets are held back because it is felt they are not ready because they are not of sufficiently high quality, and that resources prevent their speedy improvement. But data users say that lower quality is not as much of a problem as is non-publishing.

iv) This will require measured and incremental improvement. Therefore, government should commit to reporting annually on the progress that has been made to meet this twin-track policy. There should be a co-ordinated programme of audit for each department and public sector funded body of their open data performance with recommendations for further release. The system of departmental information asset registers should be standardised to make searching and navigation easier and should be expanded to include routine consideration of the suitability for publication of both structured and unstructured information.

Recommendation 3

There should be clear leadership for driving the implementation of the National Data Strategy throughout the public sector. There are many committees, boards, overseers and champions of data; but no easily understood, easily accessed, influential mechanism for making things happen. There should be a single body with a single public interface for driving increased access to PSI.

Supporting the leadership should be a “data intelligence and innovation group” to provide external challenge and aid delivery. This group, which may be linked to the ODI, should perform a non-executive role.

Detail: A review of current governance structures for PSI is needed to identify a primary channel to lead the implementation of the National Data Strategy, and the controls it can use to be most effective. This should be a simplification process, not an increase of governance complexity and it should increase the connectivity between boards/groups to limit duplication of effort and actions that are not aligned appropriately.

References

Individual Chapters:

Foreword

Thomas / Walport Data Sharing Review;
http://www.ico.org.uk/upload/documents/pressreleases/2008/thomas_walport_statement.pdf

Science as an open enterprise, Royal Society 2012;
http://royalsociety.org/policy/projects/science-public-enterprise/report/

Open Data White Paper –Unleashing the Potential
http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf

Information Governance Review – Caldicott Review
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf

European PSI Scoreboard
http://epsiplatform.eu/content/european-psi-scoreboard

Economic and Social Research Council's (ESRC) Secure Data Service
http://www.esrc.ac.uk/funding-and-guidance/tools-and-resources/research-resources/data-services/sds.aspx

Chapter 1 - Evidence

Results of the online consultation of stakeholders "Review of the PSI Directive";
http://ec.europa.eu/information_society/policy/psi/docs/consultations/cons2010/results_online_consultation_final.doc

Administrative Data Taskforce - Improving Access for Research and Policy ;
http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative-initiatives/Administrative-Data-Taskforce.aspx

Open data Dialogue - Final Report;
http://www.rcuk.ac.uk/documents/documents/TNSBMRBRCUKOpendatareport.pdf

Analysis of 'Open Data' survey commissioned to support the Shakespeare Review into Public Sector Information;
http://research.yougov.co.uk/white-papers/open-data-survey-into-public-sector-information/

Valuation Techniques for Social Cost-Benefit Analysis;
http://www.hm-treasury.gov.uk/d/green_book_valuationtechniques_250711.pdf

YouGov, undertook a survey of public opinion between 22 February and 15 March 2013 about PSI
http://datahub.io/dataset/shakespeare-review

Chapter 2 - Ownership

OS OpenData - driving growth and innovation;
www.ordnancesurvey.co.uk/oswebsite/products/os-opendata.html

Free Our Data: Make taxpayers' data available to the public;
www.freeourdata.org.uk/

The Green Book: appraisal and evaluation in central governent;
www.hm-treasury.gov.uk/data_greenbook_index.htm

Chapter 3 - Privacy

Independent transparency and privacy review;
https://www.gov.uk/government/publications/independent-transparency-and-privacy-review

Virtual Microdata Laboratory;
http://www.ons.gov.uk/ons/about-ons/who-we-are/services/vml/index.html

The HMRC Datalab;
http://www.hmrc.gov.uk/datalab/

Fair Data 10 core principles;
http://www.fairdata.org.uk/10-principles/

World Economic Forum's recent report 'Rethinking Personal Data';
http://www3.weforum.org/docs/WEF_IT_RethinkingPersonalData_Report_2012.pdf

Enforcement of the Data Protection Act 1998; Breaches of section 55 DPA;
http://www.publications.parliament.uk/pa/cm201213/cmselect/cmjust/962/96205.htm#a8

Chapter 4 - Capability

Apps for Good; Technolofy Teaching Apps;
http://www.appsforgood.org/

A nationwide network of volunteer-led after school coding clubs for children aged 9-11;
http://www.codeclub.org.uk/

GirlsWhoCode works to educate and equip young women with the skills to pursue academic and career opportunities in computing fields;
http://www.girlswhocode.com/

The Open Data Institute accredited qualification in Open Data Technologies;
http://www.theodi.org/excellence/pg_certificate

getstats is a campaign by the Royal Statistical Society to improve how people handle numbers;
http://www.getstats.org.uk/

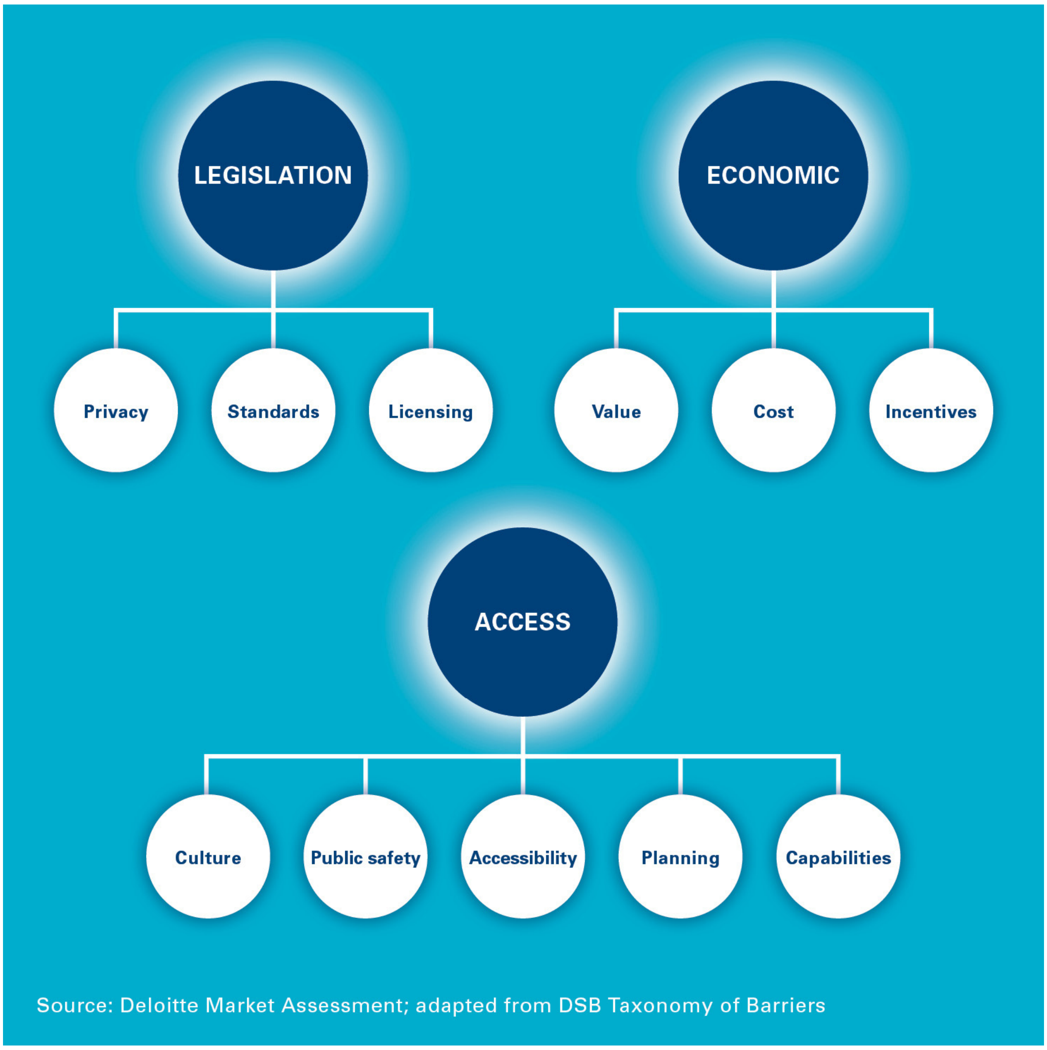
The Nuffield Foundation Quantitative Methods Programme;
http://www.nuffieldfoundation.org/quantitative-methods-programme

Cabinet Office: Meeting the challenge of change – A capabilities plan for the Civil Service;
http://engage.cabinetoffice.gov.uk/capabilities-plan/

TIMSS 2011 International Results in Mathematics
http://timss.bc.edu/timss2011/international-results-mathematics.html

Appendices

Figure 1.1: Taxonomy of Barriers



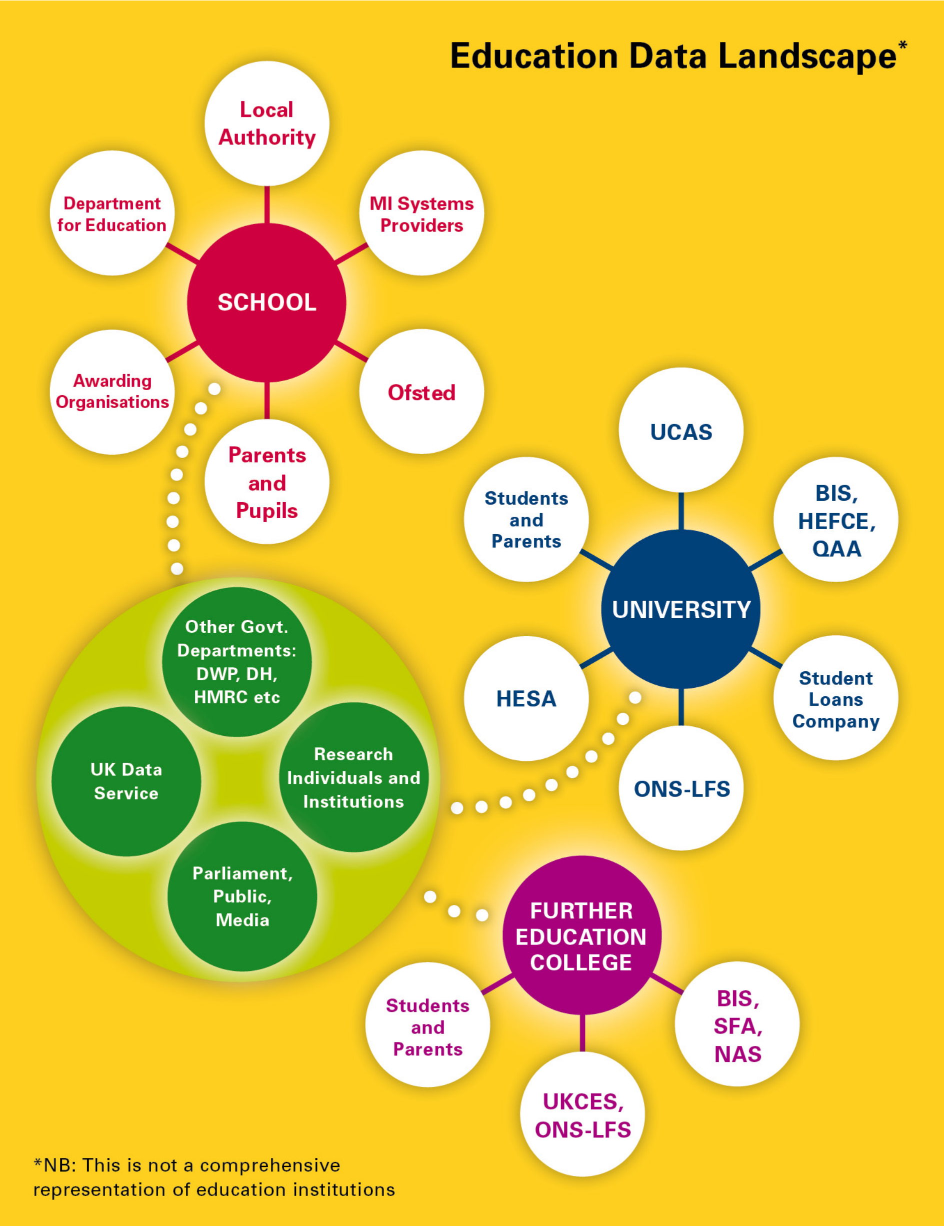
Source: Deloitte Market Assessment; adapted from DSB Taxonomy of Barriers

Appendix 1.2: Education Case Study

As part of the review two areas are considered in detailed case studies. The first of these is education data.

A range of education institutions handle and use data. Education institutions have a range of key stakeholders that produce and use data, illustrated by the diagram below.

Figure 1.2: Key education institutions and links (Source: In-house analysis)



*NB: This is not a comprehensive representation of education institutions

What data is collected and released?

The collection of data varies in light of who collects the data and what the collected data refers to.

Source of collection – who collects the data

For early years and schools data, initial collection takes place in schools and early years settings, through awarding organisations, by Local Authorities, by the Department for Education (DfE), and by the Office for Standards in Education, Children's Services and Skills (Ofsted).

Further Education (FE) data is collected by both public and private FE providers, by students, employers (especially those part-funding FE training) and by those administering and managing performance through the Department for Business, Innovation and Skills (BIS), the Skills Funding Agency (SFA) and the National Apprenticeship Service (NAS). Beyond these, the UK Commission for Employment and Skills (UKCES) and the Office for National Statistics (ONS) collect data on skill levels and associated variables.

In Higher Education (HE), the Higher Education Statistics Agency (HESA) is the official agency for the collection, analysis and dissemination of quantitative information about higher education. Universities themselves, UCAS, and the Student Loans Company also hold a vast amount of data. In addition, other public bodies - the Department for Business, Innovation and Skills (BIS), the Higher Education Funding Council for England (HEFCE), and the Office for National Statistics (ONS) – also collect HE data.

Across the range of these services, data on related variables – income levels, labour market outcomes, international student numbers, medical student numbers – will also be collected by other government departments and public bodies, examples including HMRC, the Department for Work and Pensions (DWP), UK Border Agency (UKBA) and the NHS.

Type of data – what the collected data refer to

The type of data collected in education varies along the *level* to which it refers and the *part of the education delivery chain* that the data refer to.

For *level*, data could refer to individual-level, classroom, institution-specific, regional (e.g. Local Authority), or national information. Data could equally record information across different parts of the *education delivery chain* – from inputs (e.g. resource and capital investment) and activity (e.g. number of lessons delivered) to outputs (e.g. number of graduating students) and outcomes (e.g. labour market outcomes of students). These distinctions of type – both in *level* and *delivery chain part* terms - are useful in explaining both *extent of release* and the *uses made* of data. For example, individual-level data, particularly where it relates to outcomes, will be sensitive and is likely to be limited in *extent of release* and reserved for use in improving internal teaching aimed at individuals, as well for research into learning and outcomes. National, input data, such as about funding for school buildings, will often be released publicly as a means of providing use through accountability.

The conditions under which data is released and constraints on who can access data vary across datasets. Generally, there are three fronts on which conditions may apply:

i) On *who* can access data: For example, In FE, BIS and the Skills Funding Agency publish its data in the lowest aggregated format possible, normally at provider or local authority level. In addition, FE shares personal data with third parties for research and other purposes under terms and conditions, adhering to the Data Protection Act.

ii) On *what* can be done with datasets: An example of restriction on purpose of data-use is the National Pupil Database (NPD). Data from the National Pupil Database can be shared with named bodies and persons conducting "research into the educational achievements of pupils"[57].

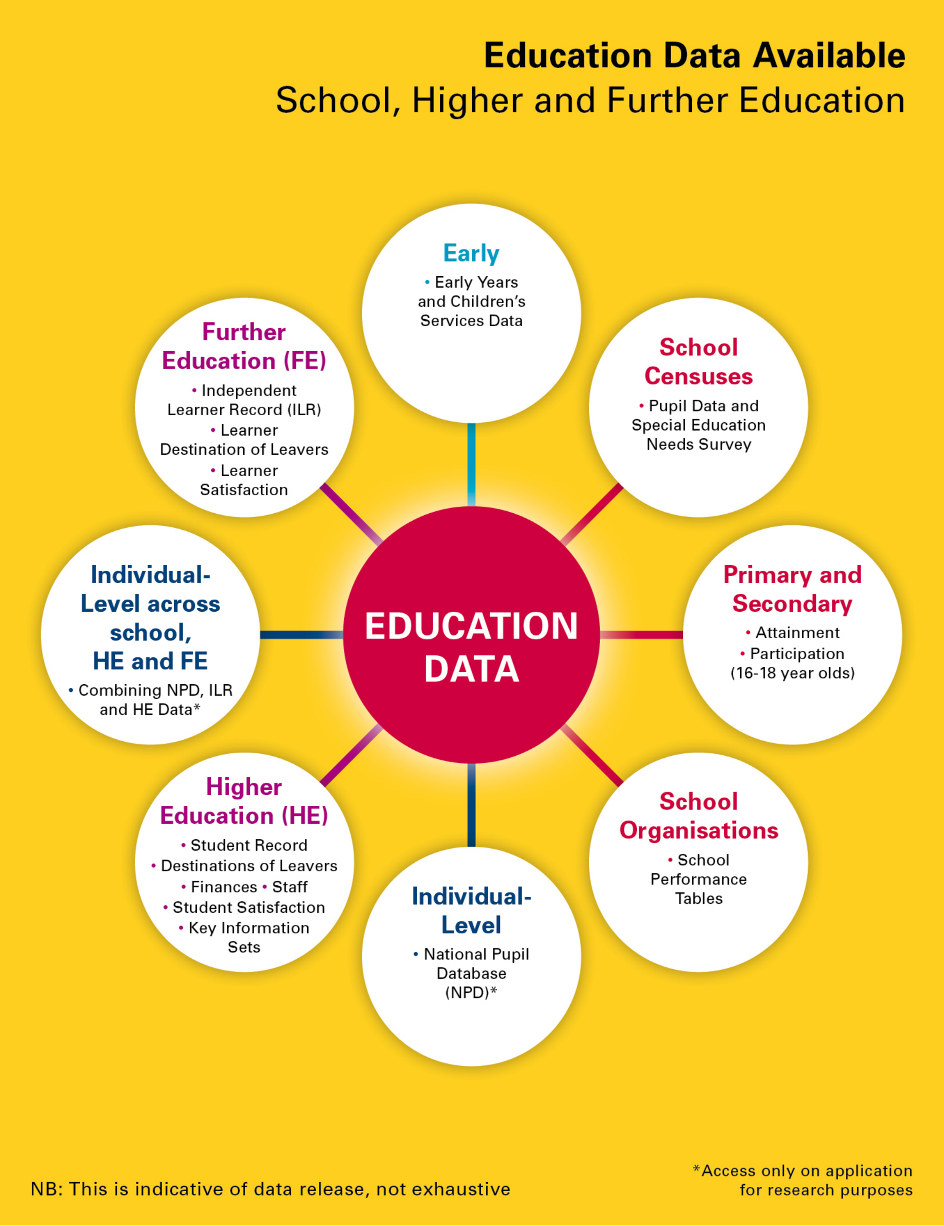
iii) On *format* of release and user interface: Some datasets are presented through interface tools which allow easy comparison by users – e.g. the revised Unistats website for HE applicants which provides the new Key Information Set data; the DfE Performance Tables Website; and FE Choices, which allows learners to compare Further Education colleges.[58] Other datasets – such as those which sit behind Department for Education Statistical First Releases (SFRs) are provided in Excel spreadsheets. Some of these datasets are available in reusable formats and can be accessed with an API – an example is data on Average Class Sizes.

Figure 2 overleaf illustrates some of the key datasets that are released, either publicly or under constraints on what can be done with the datasets.

[57] Under the *Education (Individual Pupil Information) (Prescribed Persons) (England) Regulations 2009.*

[58] On http://fechoices.skillsfundingagency.bis.gov.uk/Pages/home.aspx.

Figure 1.3: Key education data (Source: in-house analysis)



Education Data Available
School, Higher and Further Education

EDUCATION DATA

Early
- Early Years and Children's Services Data

School Censuses
- Pupil Data and Special Education Needs Survey

Primary and Secondary
- Attainment
- Participation (16-18 year olds)

School Organisations
- School Performance Tables

Individual-Level
- National Pupil Database (NPD)*

Higher Education (HE)
- Student Record
- Destinations of Leavers
- Finances
- Staff
- Student Satisfaction
- Key Information Sets

Individual-Level across school, HE and FE
- Combining NPD, ILR and HE Data*

Further Education (FE)
- Independent Learner Record (ILR)
- Learner Destination of Leavers
- Learner Satisfaction

NB: This is indicative of data release, not exhaustive

*Access only on application for research purposes

How do we use education data in the UK?

1) **Accountability**: School performance data is used to hold schools to account.

2) **Research**: The National Pupil Database is an important resource for research. It is a longitudinal database holding information on children in schools in England. There are a range of data sources in the National Pupil Database providing information about children's characteristics and education at different stages (pre-school, primary, secondary and further education). The data is used extensively to support research and present insights into attainment and provision of education services.

3) **Service improvement**: through better design, delivery, evaluation and through user choice. User choice is facilitated through key interface tools:

- Ofsted Data Dashboard: http://dashboard.ofsted.gov.uk/
- KIS University data on Unistats: http://unistats.direct.gov.uk/
- FE Choices: http://fechoices.skillsfundingagency.bis.gov.uk/Pages/home.aspx

4) **Service regulation**: Data from Further Education colleges are used to allocate funding. The FE data collected is used by organisations in the FE and Skills sector (including providers themselves) to ensure that public money is being spent in line with government targets for quality and value-for-money, for future planning and to make the case for the sector in seeking further funding.

5) **Policy design**: The National Pupil Database allows comparison across different pupil cohorts and allows tracking of progress across time (at key stage moments) for pupils. For example, it is possible to track students at Key Stage 5 (aged 18) back to their Key Stage 1 teacher assessments (aged 7). ILR and HESA data are also used to track progress.

What are other countries doing?

Programme for International Student Assessment (PISA)

PISA is a three-yearly survey of the knowledge and skills of 15-year-olds in a number of industrialised countries. Key indicators assess literacy in reading, mathematics and science, data gathered through school tests, and the assessment instruments are developed by international education experts.[59]

The USA: Education Data Initiative[60]

The Education Data initiative is run by the US Department of Education, in partnership with the White House. The Initiative seeks to expand education-related datasets that are available in machine-readable form, while protecting personal privacy. It also seeks to encourage private-sector awareness of datasets for use in subsequent product development and innovation. There is also an attempt to enable students to access electronically their assessment data, to allow students to create online learning profiles and access appropriate online learning tools.

There are four key focus projects for the Education Data Initiative:

The **MyData Initiative** seeks for every student (or parent) to have access to his or her own academic data in both machine-readable and human-readable format.

[59] *OECD Handbook for Internationally Comparative Education Statistics: Concepts, Standards, Definitions and Classifications, 2004*

[60] http://www.ed.gov/edblogs/technology/education-data-initiative/

The **Learning Registry** is a new way to identify and find educational resources online. Teachers and internet browsers can add content to the registry and tag it on the basis of its quality or key themes. Those resources which are highlighted as being of good quality emerge at the top of the registry, allowing teachers easy access to effective learning tools.

Open Badges allows colleges and industry organisations to award micro-credentials (badges) to students who demonstrate proficiency in specific competencies accessed in learning tools across a range of learning courses and tools. Because the technology behind the badges is open, a learner can collect badges from any number of different organizations and showcase them in one single place. Eventually, employers can use Open Badges to filter candidates on the basis of required skills.

The Education Data Initiative is looking to expand the extent of **raw datasets** that are released. Raw data sets are defined as public data without any personally identifiable information, made freely available to the public for download, re-use, and even commercial use.

In addition to the Education Data Initiative, the US Department of Education has a publicly accessible, regular education dashboard that shows national progress on key indicators set out by the Obama administration on education.

Other examples:

- Kenya Find My School: http://findmyschool.co.ke/
- World Bank comparative aggregative figures: http://data.worldbank.org/indicator/SE.PRM.ENRR

Table 1.1: Summary of key education data release from central government departments and other organisations

Category of data	Description of data	Frequency
Education	Opening up access to the National Pupil Database	Ongoing
	Enhanced School Performance Tables	Annual
	Destination measures	Annual
	Children in Care and Adoption Performance Tables	Bi-annual
	School Workforce Census	Annual
Further Education	FE Choices: For comparing the performance of further education colleges and other organisations that receive Government funding	Annual

	to educate and train people over the age of 16.	
	National Careers Service: To provide careers advice and information on a wide range of jobs, training course resources and funding	Continuously
	Statistical First Release (SFR) & its significant cascade covering including local authority and provider level data	4 times a year
	Additionally, some operational data such as the Education and Training National Success Rates Tables (NSRT) and Skills Funding Agency Provider Allocations data are released.	Annual
Higher Education	Key Information Set: this includes extracts from the National Student Survey and Destination of Leavers (DLHE) Surveys, as well as other data supplied from HEIs to HEFCE, who collate and then distribute the KIS.	annual (introduced Sept 12)
	HEFCE – National Student Survey	annual
	HESA - Statistical First Release - Destinations of Leavers from Higher Education in the United Kingdom	annual
	HESA Performance Indicators	annual
	HESA - Statistical First Release - Higher Education student enrolments and qualifications obtained at Higher Education Institutions	annual
	HESA – Staff in HEIs	annual
	HEIPR	annual

Appendix 1.3: Health Case Study

As part of the review two areas are considered in detailed case studies. The second of these is healthcare data.

What data is collected and released?

Internationally, the UK has collected one of the most comprehensive sets of national healthcare data.[61] The UK has benefitted from a single National Health Service (NHS) collecting data for decades.[62] At a central level, bodies such as the Department of Health (DH) collect a range of data for national management and published statistics.

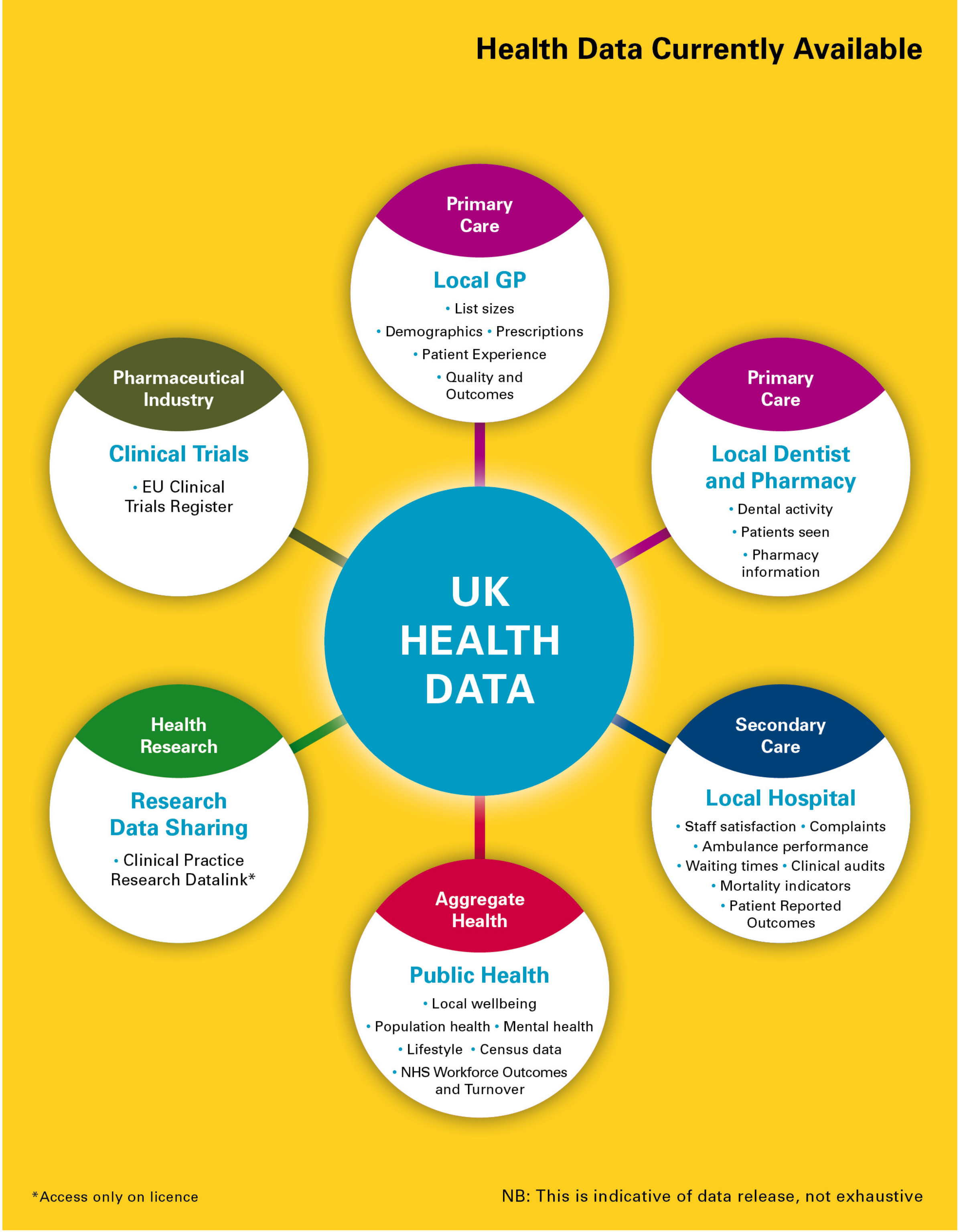
Much NHS data is used at local level. In addition, with the transfer of public health to local authorities from 1 April 2013, there is a significant amount of work going on to better integrate health and social care, and a substantial amount of broader health-related data held by Councils.

Overleaf the infographic maps some of the key datasets released to the general public or, in the case of the Clinical Practice Research Datalink (CPRD), to researchers on licence.

[61] OECD (2013). *Exploring data-driven innovation as a new source of growth: mapping the policy issues raised by 'Big Data'*.

[62] McKinsey Global Institute (2011). *Big data: the next frontier for innovation, competition and productivity*, McKinsey & Company.

Figure 1.4: Current healthcare data (source: in-house analysis)



Health Data Currently Available

UK HEALTH DATA

Primary Care

Local GP

- List sizes
- Demographics • Prescriptions
- Patient Experience
- Quality and Outcomes

Primary Care

Local Dentist and Pharmacy

- Dental activity
- Patients seen
- Pharmacy information

Secondary Care

Local Hospital

- Staff satisfaction • Complaints
- Ambulance performance
- Waiting times • Clinical audits
- Mortality indicators
- Patient Reported Outcomes

Aggregate Health

Public Health

- Local wellbeing
- Population health • Mental health
- Lifestyle • Census data
- NHS Workforce Outcomes and Turnover

Health Research

Research Data Sharing

- Clinical Practice Research Datalink*

Pharmaceutical Industry

Clinical Trials

- EU Clinical Trials Register

*Access only on licence

NB: This is indicative of data release, not exhaustive

How do we use healthcare data in the UK?

Healthcare data are extensively used in the UK. NHS data is used by hospitals, linked medical data used by researchers and pharmaceutical companies, data on local services used by signposting services, personal health data used by individuals to shape behaviour. In a majority of cases, individual uses of data do not draw on government open data or other data owned by government bodies (for example, personal use is mostly centred on data owned and analysed by the individuals it relates to), not all uses of data owned by public bodies will be based on open healthcare data (for example, many uses of healthcare data take place within the NHS, with hospitals drawing on their own information in more innovative ways to drive efficiencies, as in the University Hospital Birmingham case study below).

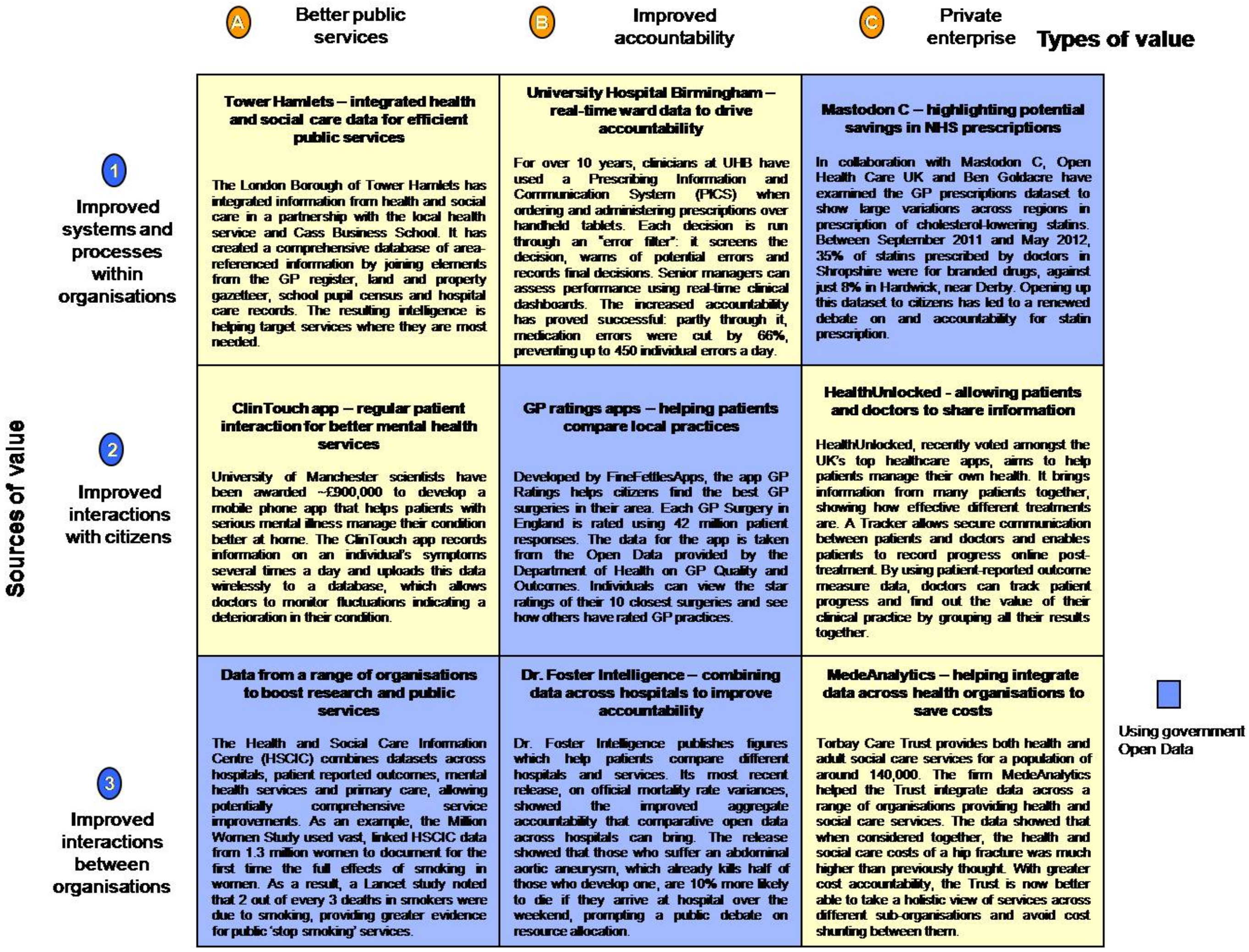
Some of the key uses in healthcare have benefitted from the ***volume***[63] of big data. The National Institute for Health and Clinical Excellence (NICE), for example, has used large clinical databases to examine the cost effectiveness of new drugs and treatments. The UK Life Sciences Strategy[64] mentions a number of uses of vast datasets, including a project to sequence 100,000 whole genomes at diagnostic quality, to pursue pioneering developments in the life sciences.

In other instances, prompt service delivery has been possible as a result of the ***velocity*** of real-time data. Some of these uses, e.g. the use of real-time information to better manage at University Hospital Birmingham[65], have drawn on the latest developments in technology to facilitate the timely collection and transmission of practitioner and patient data. Others do not require real-time information, but benefit from relative increases in pace: e.g. a consistent, timely record of complaints.

Finally, there are examples of data usage which make use of the ***variety*** of available data, linking hitherto disparate datasets to generate key healthcare insights. A good example is that of Torbay Care Trust, which integrated data across a large number of health and social care organisations for a holistic view of and control over costs[66].

[63] OECD, *Exploring data-driven innovation as a new source of growth*, DSTI/ICCP(2012)9
[64] HM Government (2012), *Strategy for UK Life Sciences: one year on.*
[65] University Hospital Birmingham – Reform (2012), *Doctors and nurses*
[66] See http://www.microsoft.com/health/en-gb/articles/Pages/torbay-care-trust.aspx

Figure 1.4: Case studies of healthcare data usage[67] (source: drawing on BCG analysis)



Sources of value / Types of value	A Better public services	B Improved accountability	C Private enterprise
1 Improved systems and processes within organisations	**Tower Hamlets – integrated health and social care data for efficient public services** The London Borough of Tower Hamlets has integrated information from health and social care in a partnership with the local health service and Cass Business School. It has created a comprehensive database of area-referenced information by joining elements from the GP register, land and property gazetteer, school pupil census and hospital care records. The resulting intelligence is helping target services where they are most needed.	**University Hospital Birmingham – real-time ward data to drive accountability** For over 10 years, clinicians at UHB have used a Prescribing Information and Communication System (PICS) when ordering and administering prescriptions over handheld tablets. Each decision is run through an "error filter": it screens the decision, warns of potential errors and records final decisions. Senior managers can assess performance using real-time clinical dashboards. The increased accountability has proved successful: partly through it, medication errors were cut by 66%, preventing up to 450 individual errors a day.	**Mastodon C – highlighting potential savings in NHS prescriptions** In collaboration with Mastodon C, Open Health Care UK and Ben Goldacre have examined the GP prescriptions dataset to show large variations across regions in prescription of cholesterol-lowering statins. Between September 2011 and May 2012, 35% of statins prescribed by doctors in Shropshire were for branded drugs, against just 8% in Hardwick, near Derby. Opening up this dataset to citizens has led to a renewed debate on and accountability for statin prescription.
2 Improved interactions with citizens	**ClinTouch app – regular patient interaction for better mental health services** University of Manchester scientists have been awarded ~£900,000 to develop a mobile phone app that helps patients with serious mental illness manage their condition better at home. The ClinTouch app records information on an individual's symptoms several times a day and uploads this data wirelessly to a database, which allows doctors to monitor fluctuations indicating a deterioration in their condition.	**GP ratings apps – helping patients compare local practices** Developed by FineFettlesApps, the app GP Ratings helps citizens find the best GP surgeries in their area. Each GP Surgery in England is rated using 42 million patient responses. The data for the app is taken from the Open Data provided by the Department of Health on GP Quality and Outcomes. Individuals can view the star ratings of their 10 closest surgeries and see how others have rated GP practices.	**HealthUnlocked - allowing patients and doctors to share information** HealthUnlocked, recently voted amongst the UK's top healthcare apps, aims to help patients manage their own health. It brings information from many patients together, showing how effective different treatments are. A Tracker allows secure communication between patients and doctors and enables patients to record progress online post-treatment. By using patient-reported outcome measure data, doctors can track patient progress and find out the value of their clinical practice by grouping all their results together.
3 Improved interactions between organisations	**Data from a range of organisations to boost research and public services** The Health and Social Care Information Centre (HSCIC) combines datasets across hospitals, patient reported outcomes, mental health services and primary care, allowing potentially comprehensive service improvements. As an example, the Million Women Study used vast, linked HSCIC data from 1.3 million women to document for the first time the full effects of smoking in women. As a result, a Lancet study noted that 2 out of every 3 deaths in smokers were due to smoking, providing greater evidence for public 'stop smoking' services.	**Dr. Foster Intelligence – combining data across hospitals to improve accountability** Dr. Foster Intelligence publishes figures which help patients compare different hospitals and services. Its most recent release, on official mortality rate variances, showed the improved aggregate accountability that comparative open data across hospitals can bring. The release showed that those who suffer an abdominal aortic aneurysm, which already kills half of those who develop one, are 10% more likely to die if they arrive at hospital over the weekend, prompting a public debate on resource allocation.	**MedeAnalytics – helping integrate data across health organisations to save costs** Torbay Care Trust provides both health and adult social care services for a population of around 140,000. The firm MedeAnalytics helped the Trust integrate data across a range of organisations providing health and social care services. The data showed that when considered together, the health and social care costs of a hip fracture was much higher than previously thought. With greater cost accountability, the Trust is now better able to take a holistic view of services across different sub-organisations and avoid cost shunting between them.

🟦 Using government Open Data

Potential obstacles and costs

The release of healthcare data, especially good quality data, imposes costs - large fixed costs in acquiring IT systems and potential variable costs in regularly providing quality assurance and supervision of collected data. Making the business case for continued release of data relies on making robust cases of costs and benefits, as well as of the timeframe over which benefits accrue.

[67] Sources: Tower Hamlets case study – Department of Health (2012). *The power of information: putting us all in control of the health and care information we need*; University Hospital Birmingham – Reform (2012), *Doctors and nursess*; Mastodon C – The Economist (2012), '*Open data and healthcare*'; ClinTouch app – HM Government (2012), *Strategy for UK Life Sciences: one year on.*; GP ratings apps – Department of Health website; HealthUnlocked – Department of Health website GP apps ratings; HSCIC – HM Government (2012), *Strategy for UK Life Sciences: one year on.*; Dr. Foster Intelligence – Dr. Foster Intelligence website; MedeAnalytics – OECD (2013). *Exploring data-driven innovation as a new source of growth: mapping the policy issues raised by 'Big Data'*.

With any use of health data, ensuring the privacy of patients is a central concern. Particularly as data becomes more open, datasets bigger and different datasets more linked, it becomes harder to ensure that no feasible combination of datasets will impinge upon individuals' privacy concerns by giving away information that can be de-aggregated. In light of these challenges, an Information Governance Review led by Dame Fiona Caldicott has recently been published, which looks at appropriate ways of balancing privacy concerns and the need to improve patient care in health and social care. [68]

Healthcare data is increasingly held across sectors – public NHS organisations as well as private individuals and providers. In that context, encouraging data to be shared to derive maximum value requires an intellectual property rights ownership model that aligns private interests (e.g. in privacy and commercial sensitivity) with social interest in generating collaborative uses of data.

What are other countries doing?

Compared to many other counties, the UK benefits from a less fragmented healthcare provider landscape, and therefore more comprehensive datasets. Nevertheless, there are several innovative examples elsewhere of healthcare data being used to derive value within and outside healthcare services.

The examples below – from Finland, Japan, the USA and the World Bank – show further ways in which clinical and public health data are being collected, distributed and used for better services, advanced research and increased accountability.

Finland – linked individual-level data for accurate clinical performance management and evidence-based policy[69]

In Finland, the Performance, Effective and Cost of Treatment Episodes (PERFECT) programme monitors the content, quality and cost-effectiveness of treatments for selected disease groups and procedures (e.g. stroke, premature newborns, hip fracture, breast cancer, schizophrenia), selected on grounds of prevalence and costs. By linking a range of individual-level data - hospital in-patient records, out-patient records, birth records, disease-specific registers, prescribed medicines data, social care data, death records and data on care reimbursement - PERFECT is able to monitor the whole care cycle for well defined patient groups. Generated data can be used for performance management and for benchmarking clinical practices. The data also highlight regional variations which may have arisen from varying policy decisions, providing a greater evidence base for subsequent policy decisions.

Japan – *My Hospital Everywhere* to provide patients access to electronic health records[70]

Japan's *My Hospital Everywhere* project allows patients to store and access their electronic medical records and enables doctors to access these in organisations across the country. Individuals obtain their information from the medical institutions that

[68] https://www.gov.uk/government/publications/the-information-governance-review

[69] OECD (2013). *Exploring data-driven innovation as a new source of growth: mapping the policy issues raised by 'Big Data'.*

[70] Japanese Government (2010), *A New Strategy in Information and Communications Technology (IT)*

voluntarily provide information via mobile phones and over the internet. Individuals can manage their own medical information and can present it on tablet terminals and PCs at any medical service provider they access, as well as using the records for personal health management. The project is seen as a potential route to quicker targeting of appropriate medical services and greater patient control over health.

USA - MIT Media Lab project to collaboratively map disease outbreaks in real time[71]

The MIT Media Lab has developed a range of products which draw on healthcare data to generate research and service innovation. One product of the project is Outbreaks Near Me, an app that allows people to report activity indicative of an outbreak. No report means anything by itself, but taken as a whole the data might provide timely indicators of outbreaks, far earlier than healthcare organizations would record them, delivering real-time intelligence for a range of private and public audiences.

World Bank – Comprehensive HealthStats database[72]

HealthStats is the World Bank's comprehensive database of Health, Nutrition and Population (HNP) statistics. It includes over 250 indicators covering, amongst others, health financing, HIV/AIDS, immunization, malaria and tuberculosis, water and sanitation data. Users can access HNP data by country, topic, or indicator, and view the resulting data (and wealth quintiles) in tables, charts or maps that can be easily shared through email, Facebook and Twitter. The World Bank DataFinder Health app also provides comparative visualisations and current data from the World Bank API on phones.

Further examples of healthcare data distribution and usage

- A small number of hospitals in the USA are using data on patients and conditions to target more accurately interventions on high-risk, high-cost patients, realising large savings.[73]

- Danish citizens are able to see all their hospital records online.[74]

Table 1.2: Summary of key healthcare data release from the Health and Social Care Information Centre and DH

Important Note: In general, central Government bodies collate only the data needed to meet requirements for national accountability or broader public use of summary data. Where this information meets the definition of statistics, existing guidelines and rules steer data owners to put the material in the public domain quickly. But it is important to note that these national datasets form only a small portion of the health and care data in circulation in the wider system.

[71] MIT Media Lab Human Dynamics Group: http://hd.media.mit.edu/
[72] World Bank Open Data Blog: http://blogs.worldbank.org/opendata/
[73] The New Yorker, accessed on http://www.newyorker.com/reporting/2011/01/24/110124fa_fact_gawande
[74] Secretary of State for Health (2013), accessed on http://mediacentre.dh.gov.uk/2013/01/16/16-january-2013-jeremy-hunt-policy-exchange-from-notepad-to-ipad-technology-and-the-nhs

Category of data	Detail of data	Frequency
Audits and performance	Complaints data by NHS hospital	Annual
	Clinical audit data on priority areas: cancer, diabetes, dementia, heart disease, kidney care	Annual
	Performance and activity data series, inc. data on waiting times, counts of some activity (e.g. outpatient appointments), winter situation reports, breaches of mixed sex accommodation policy etc.	Variable (ranges from weekly to annual)
	Quality and Outcomes Framework data for GP practices: management of common diseases, practice organisation, patient experience, extra services e.g. child health services	Monthly
	NHS Outcomes key outcomes: i) preventing premature deaths ii) enhancing quality of life for people with long-term conditions iii) helping people recover from ill health episodes iv) ensuring positive experience of care v) treating in a safe environment	Quarterly
	Ambulance statistics	Annual
Health and lifestyle	Health Survey for England and Infant Feeding Survey	N/A
	Data on: alcohol consumption, contraception, diabetes, diet, drug misuse, immunisation, NHS stop smoking services, obesity, physical activity and smoking	N/A
Hospital care	Patient Reported Measures for: hip and knee replacements, hernia and varicose veins	Aggregate – monthly By provider - quarterly
	Accident and Emergency, Cancer, Coronary heart disease, Critical care, hospital activity, maternity, outpatients and mortality indicator	N/A
Mental health	Trends in patients detained under Mental Health Act 1983, NHS mental health services information and Mental Health surveys	N/A

Population, geography and international	Vast range of population geography, lifestyle, wellbeing and public health information (indicators of population health), local basket of inequalities indicators	
Primary care	Comparative clinical outcomes of GP practices: patient experience, demographics, , quality outcomes, infrastructure and impact on NHS resources	Data refreshed as become available
	GP reference data: location, list sizes and demographics	Monthly
	GP practice profiles for cancer	Annual
	GP prescriptions data	Monthly
	Information on activity of pharmacy, dentistry services and trends in NHS funded sight tests	
Social care	Care Quality Commission's Provider Profile reports on care provider compliance	TBC
	Local Accounts: relevant information about comparative adult social care provision across councils	Annual
Workforce	NHS staff satisfaction, engagement, numbers, earnings, turnover and sickness absence rates	Annual
Education	Quality of post-graduate medical education by provider	Annual

In addition, DCLG publishes annual wellbeing measures at Local Authority level with interesting visualisation on Open Data Communities. It also includes Deprivation Index Health and Disability statistics.

BIS/13/744