

Exploiting government's information assets for the public good

SUMMARY

1. The Prime Minister has tasked Sir Tim Berners-Lee and Professor Nigel Shadbolt with making much more Public Sector Information (PSI) easily accessible over the web. In parallel, the Shareholder Executive has been studying how best to make available information from Trading Funds in a sustainable way. Both organisations have asked government's Advisory Panel on Public Sector Information (APPSI) for suggestions on how to facilitate their work. Since the two tasks interact and success will need both short and long term actions, this note sets out a consolidated view by APPSI¹ of how best to proceed.
2. Our key points are:
 - The value of PSI to citizens and businesses alike is now increasingly widely recognised. The factors which block realisation of this value are no longer technological but rather cultural, policy, institutional and financial ones. We therefore warmly welcome the *Making Public Data Public* initiative;
 - We identify areas where information 'prospecting and harvesting' would be beneficial in the short term. But we also urge that a longer term and prioritised information garnering strategy should be put in place;
 - In particular, we identify some 'Core Reference Geographies' which would underpin many activities of the state, its organisations (e.g. emergency services) and businesses and urge that these are made freely available and maintained as a key part of a national information infrastructure;
 - We identify some policy changes and other actions which would considerably facilitate successful re-use of PSI;
 - We note that the growing use of the new technologies has major ramifications for current government policies and practice applied to Trading Funds, notably in the practicability of some forms of end user licenses.

BACKGROUND

3. Changes in technology and an acceptance of the need for greater transparency on the part of government have led to a widening of access to information held by public sector bodies in many parts of the world. This also occurs at different levels: as just one example, the City of New York announced on 6 October 2009 that it had released 194 datasets -- on everything from restaurant inspections and property sales to traffic data and the locations of laundry facilities. The city also launched a \$20,000 competition to develop software applications using this data².

¹ The members of APPSI are drawn from a wide variety of backgrounds including providers, re-users and consumers of public-sector information, experts from academia and industry, and representatives of producer and consumer groups and Northern Ireland, Scotland and Wales. This paper is drawn together from submissions provided by all these individuals and circulated to members for comments and corrections but it should not be assumed that all of their views are identical.

² <http://www.technologyreview.com/communications/23617/?nlid=2410&a=f>

4. The information holdings of the many thousands of Public Sector Bodies in the UK are potentially of great value to the citizen and business community alike. Interest in making them more readily available for exploitation has grown rapidly in recent months but the overall UK situation is now complex for the following reasons:
- the Operational Efficiency Programme approach to addressing Trading Fund business models via the Shareholder Executive has been joined by the *Making Public Data Public* initiative mandated by the Prime Minister. Many other relevant initiatives (e.g. the Location Strategy and Location Council, the Information Assets Register, Data Grand Challenge) have been launched in recent years. In addition, the Trading Funds are mandated under legislation specific to them but also have to accommodate to the Information Fair Trader Scheme and Public Sector Information Re-Use Regulations. These laws, guidance, programmes and initiatives are not closely aligned;
 - experience in other countries indicates that usage levels of PSI rise sharply once PSI is made more easily accessible and if it is priced on a marginal cost basis. There is ample evidence that easy access to PSI fosters innovation in services and job creation. But wider PSI use and continuing revenues to government from information trading may be in conflict, at least in the short term. The current financial situation is leading HMT to challenge strongly any additional public expenditure to replace any reductions arising from change of policy;
 - many licences for re-use of Trading Fund information have been highly restrictive and complex – a situation which is now recognised and where attempts are being made to simplify them. But the advent of technology which facilitates ‘mashups’ – and hence extends the range and ease of use of information in combination – has major consequences for current government policy. The practical consequences of widespread dissemination of information ‘mashed up’ from government web sites and elsewhere without an end user license pose fundamental challenges to the Trading Fund licensing model hitherto operated;
 - there remains confusion about what government needs to have for its own purposes (under the ‘Public Task’ of its component bodies) and hence which falls under the Re-use of Public Sector Information Regulations 2005³ (even though making PSI available for re-use is not mandatory under the EU Re-Use Directive or those UK regulations). This confusion and resulting uncertainty impacts on commercial investment decisions, affects re-use by third sector and deters experiments with information;
 - in at least a few cases, multiple public sector bodies are in competition in providing information to customers;
 - there is much private sector concern – some manifested in submissions to the Office of Public Sector Information (OPSI) - that government bodies collecting information and trading in it⁴ are behaving inappropriately in regard to their private sector partners or competitors;
 - historical practice has largely been that what government needs for its own purposes has been collected predominantly by government bodies. Given the evolution of a globalised private sector skilled in information assembly and the advent of new technology which lowers the entry price to such businesses, it is no surprise that

³ SI 2005 No. 1515

⁴ Such as local authorities’ property search services

increasing tensions have occurred between government and commercial information collectors and assemblers.

GOVERNMENT AND PUBLIC NEEDS FOR INFORMATION

5. Government is committed to evidence-based approaches to decision-making and to ensuring that the operational performance and effectiveness of its components can be assessed. These approaches result in the need to have access to large swathes of information. Changing priorities also dictate periodic changes to the nature of and classifications used for such information.
6. In general, no central body determines what information is needed by government: that is a matter for individual government departments though a few bodies exist simply to collect, process and make available information for others (e.g. Office for National Statistics, Ordnance Survey and the Meteorological Office). Some others generate and license valuable information as a by-product of their administrative functions (e.g. DVLA and HM Land Registry). In addition, co-operative endeavours across local government have led to the National Land Information System, the National Street Gazetteer and the National Land and Property Gazetteer.
7. Beyond this need to support internal decision-making and external accountability in central government however there is a rapidly enlarging appetite on the part of members of the public to have access to some (e.g. location-specific) information (e.g. where is the nearest Job Centre? How is my nearest hospital or GP performing compared to others in the same area?).
8. That said, the extent of the demonstrable need for certain types of information is highly variable. In some cases, this is well-documented and pressing, such as in the highly detailed and country-wide information required for a National Emergency Services Gazetteer, as defined by a Working Group of the Association of Chief Police Officers responding to the mandates of the Civil Contingencies Act 2004 for emergency services to collaborate. Though many of the information components for this Gazetteer presently exist, they are held by different players, not integrated and their combined use is complicated by licensing constraints. The EU INSPIRE Directive can also be seen as a specification of (spatial) information which is required for government to carry out its multiple roles. On the other hand, it is not possible to predict all needs: information is often an 'experience good' whose benefits become clear only after experiment and experience. There is good evidence that 'bottom up' developments and innovation do occur when a range of information is made available to the wider public: new products and services arise, some of which will be highly successful. In what follows, we advocate and deal with both the planned and serendipitous approaches.
9. We see major opportunities – and some serious challenges - arising from the rapidly growing interest in PSI re-use. In particular, we see the *Making Public Data Public* initiative as providing a strong impetus to the PSI cause. We note however that there have been multiple attempts to facilitate re-use of PSI in the recent past. For example, the 2006 Data Grand Challenge set out to “highlight how Government can enable the delivery of innovative solutions by resolving issues that impede data access by:
 - Identifying existing barriers (legal, technical, social, economic or political) that impede data sharing and how these might be addressed;

- Commissioning the rapid development of novel beta applications to demonstrate how cross-government data improve the provision of public services and inform policy.”⁵
10. We recognise therefore that some ‘road blocks’ to progress have hitherto proved difficult to dislodge and some at least are unlikely to be resolved in a short time frame. So we take a holistic view of how best to proceed: our suggestions below embrace both the short and long term in proposing a battery of approaches.

THE WAY AHEAD

11. We see four complementary elements as being essential.

Celebrate and extend existing successes

12. Examples of success stories already exist. These include the exploitation by hospitals and other NHS bodies of a wide variety of information re-packaged by Dr Foster from that supplied by the NHS Information Centre; this needs to be made more widely available and easier to use. The ONS Neighbourhood Statistics and their equivalents in devolved administrations form a crucial (and free) source for many professionals and businesses but could also be expanded and made easier to use. The re-jigging of the London Gazette made available by OPSI has been little short of revolutionary. Using these as early stage exemplars, along with use of public competitions such as *Show Us A Better Way*⁶, will foster demand, provision and public take-up.

Garner additional PSI and make it immediately available

13. The Berners-Lee and Shadbolt strategy is predicated upon harvesting ‘low hanging fruit’ which is not yet readily available but which exists across a range of government departments. The approach, allied to top-level political support, could generate a ‘rolling snowball’ of newly re-useable information and should serve as a major encouragement to all departments to progress the PSI re-use agenda. Hence it is a good thing. We see advantages in the initiative focusing on central government and NHS data in the first instance since there are already mechanisms in place for harmonising the information (the situation in regard to LA data is somewhat more complex). We also see benefits in including information from all four administrations in the UK though we recognise that some information can not be compared directly (e.g. the four Indexes of Multiple Deprivation)
14. APPSI colleagues are able from their individual experience to suggest examples of apparent ‘low hanging fruit’. For example, the NHS Business Services Authority has a Prescriptions Pricing Division (PPD) whose data we believe is worth about £50m annually to the pharmaceutical companies alone and which, released and merged with other data, could provide just the sort of stimulus to the healthcare information industry provided by Dr Foster. Hitherto, however, the PPD has resisted applications for this data’s re-use, due to the GP’s lobby which seems to believe it can be used to trace individual GP prescribing behaviours. It is of course simple to ensure that the data can

⁵ <http://www.dft.gov.uk/pgr/scienceresearch/csa/datagrandchallenge>

⁶ We note that the Australian government has just followed the example of the UK government in running a similar competition. The Australian Government 2.0 Taskforce (<http://gov2.net.au/>) has announced it’s MashupAustralia (<http://mashupaustralia.org/>) contest, offering more than AU\$20,000 in prize money for the best mashups of Australian government data. As part of the initiative, more than 59 datasets from 15 different government bodies - from Federal electoral boundaries to the locations of public toilets - have been released under Creative Commons licences. The government of Denmark has also announced its intention to run such a competition (and see footnote 2).

have the postal codes stripped out or aggregated to a higher level. Though we think this information could with benefit be made available, we do not under-estimate the negotiation effort involved.

15. There are several sources of 'desired information' held by government. The *Unlocking Service* of OPSI is a good example: recent, well-argued requests on that web site include desirable enhancements to access for patent data, school Ofsted reports, Companies House basic data and information on government overseas aid. The Guardian's *Free our Data* campaign has identified multiple sources of PSI over the last three years which it claims are restricted in re-use by technical or administrative barriers. Government departments, notably the Department for Transport, have supported web-based explorations of their information (e.g. commuting time maps produced by MySpace for DfT) which would benefit from that information being made routinely and easily available. Furthermore, each government department is required to produce and maintain an Information Asset Register. These are fruitful sources of potentially valuable information if made more generally available. Unfortunately, not all areas of government produce such registers, many of those that do exist are not being maintained up-to-date and others have patchy coverage⁷. In some cases these registers are also difficult to interpret: indeed the entire collection of Information Asset Registers could be improved by use of semantic web technology.
16. Quite apart from numerical information, there is much to be gained from easing access to much textual information and, in particular, exploiting it via such semantic web technologies. For example, the results from the £1.8bn research work funded annually by the seven research councils are described briefly on their web sites but access to these is via traditional keyword search approaches and requires real expertise and perseverance to identify results which might be relevant to a commercial organisation. The extensive holdings – some in digital but much still in paper form - in the British Library's national newspaper collection at Colindale could be transformed in value by new technology more sophisticated than the tools currently available (results of about a third of all searches are now returned in PDF form).
17. In addition to 'liberating' new PSI, it would be helpful to fix those types of problems with 'already available information' whose utility is compromised by the form in which they are made available (e.g. Environment Agency pollutions prosecutions information which is public domain information: as supplied in PDF form, it can not readily be linked with (e.g.) Dunn and Bradstreet company information).
18. In seeking to make available many more sets of information, Sir Tim and Professor Shadbolt could also make a highly valuable contribution by identifying the quality of the data and by having regard to the difficulty of maintaining anonymisation when it is exploited via new profiling and mining tools. If not tackled, such factors could reduce the value of the 'liberated information' to much lower levels.
19. Finally, information alone is not enough. Making available the crawler and semantic web technologies developed by Professor Shadbolt's own department at the University of Southampton or equivalent tools would enable public bodies and their suppliers to *Make Public Data Public*. Many departments facing resource cuts might well be more responsive if standard templates and other aids to describing and up-loading information (where necessary) were available. Whilst there are some government bodies which have already embraced these new approaches (such as Ordnance Survey and the

⁷ The new requirements relating to the responsibilities of Senior Information Risk Owner may prove helpful.

Office for National Statistics), fostering awareness of them more widely across government is also highly desirable.

Follow a longer term prioritised information harvesting strategy

20. The dangers of relying solely on any initiative are two-fold: that it becomes a ‘shot gun’ approach which is not necessarily efficient since some existing sets of information are core to many more applications than others and it may expire before some fundamental and difficult issues (e.g. aspects of licensing) have been resolved. So we see a long term strategy as complementing the *Making Public Data Public* initiative.
21. **Components of a national information infrastructure.** The extent to which we have identified core sets of information needed across government and beyond is variable but some needs at least are already clear. We see the most important of these as forming a national information infrastructure. Roper and Barr⁸ have argued that certain apparently mundane ‘Core Reference Geographies’ form a vital ‘spatial element’ of this infrastructure which is required to run a country in the early twenty first century. They have argued that their Core Reference Geographies are directly analogous to the Domain Naming Service (DNS) for the web – without which it could not function. DNS provides a link between a web or email address and the appropriate numerical address of a piece of hardware. It is maintained in real time through a simple regulatory framework and global standards and its use is free to Internet users at the point of use. In particular, these geographies were said by Barr and Roper to be:
- A truly (UK) national address, street⁹ and postcode file, assembled to meet a range of known purposes from existing components currently produced by several public sector bodies and maintained in perpetuity. This is core to a range of operational activities (such as routing of ambulances or other emergency services). Most of the components of this already exist in various organisations but need to be brought together. As just one example of the present situation, some £12m is having to be expended in the next Census of Population to bring together a more comprehensive but ‘one off’ set of address information from the existing sources; the Public Administration Select Committee has strongly criticised the current situation;
 - The boundaries of administrative areas (such as Wards, Parliamentary constituencies, local authority, NHS and national boundaries) – plus areas such as Sites of Special Scientific Interest, flood zones, school catchment areas and other planning zones and others for which much information is already freely available (such as Output Areas for Census information). Boundaries are needed to be able to relate properties to administrative areas e.g. for taxation purposes.
22. Many other knowledgeable individuals have argued that some forms of mapping also form part of a national information infrastructure. There is, for example, a positive externality derived from everyone using the same ‘geographical framework’; this ensures that any information fitted to that framework will itself fit together with all other information fitted to it. In the absence of more detailed and quantitative studies as to the level of detail required, it has been plausibly argued that wide-area needs of citizens and many national government interests can mostly be met by mapping equivalent to 1:10,000 scale i.e. the most detailed forms of Ordnance Survey information and those of

⁸ <http://www.agi.org.uk/SITE/UPLOAD/DOCUMENT/Events/AGI2009/papers/RobertBarr.pdf>

⁹ We believe this must include all routes between places

its commercial and social enterprise competitors might not be included in this categorisation¹⁰.

23. The essence of success is that these information sets comprising the national spatial information infrastructure should be up-to-date, be widely and easily accessible, available at marginal or zero cost for any purpose and be national in extent. We recognise that acceptance of the Barr and Roper proposal would entail some additional public expenditure in disentangling the complex of intellectual property rights at present and maintaining the information up-to-date – but arguably would bring very major benefits.
24. **Other information to be prioritised.** We have tried to establish a conceptual framework for prioritising other government information assets which should be made available for easy re-use. Some at least of these assets will also form part of a national information infrastructure. In simple terms, this conceptual framework could be considered as a taxonomy of people, place, organisations and rules. More concretely, it would include the following categories:
- Descriptions of the characteristics of the population and of the economy;
 - The laws of the land (where we do not yet have a fully up-to-date and comprehensive statute law database: see <http://www.law.ed.ac.uk/sln/blogentry.aspx?blogentryref=7872>);
 - Information on those whom we have elected or appointed to public office;
 - Information on what all government bodies do on our behalf;
 - Details of government expenditure incurred by governments;
 - Information to monitor the performance of governments, both absolutely against any targets they set or promises they make and relatively between comparable government bodies;
 - Information on new resources available to citizens and businesses alike as a consequence of government investments (see below).
25. Much information already exists relevant to these non-exhaustive categories (e.g. from the central statistics hub at <http://www.statistics.gov.uk/hub/index.html>), Parliamentary and unofficial sources such as <http://www.theyworkforyou.com>), HMT publications, PSA targets and achievements against them and reports from the Audit Commission and other regulators or monitoring organisations). But navigating the information landscape is often complex and huge variations of approach exist in how the information is made available. For example, we have been told that access in readily useable form to such basic sets of information as Council Tax and Business Rates would be a huge benefit to many businesses. We have already (above) indicated the way in which research outputs from £1.8bn per annum of government expenditure could be made much more useable.
26. An important recent development is the growth of many administrative data bases across government. The DWP, for example, has assembled an extraordinary database combining social security, tax and other set of records for every known individual in the country. This database includes over 65 million records. Within that department this underpins routine administration, helps to guard against fraud and provides resilience against risks of disaster (e.g. through destruction of records by fire). Aggregated to groups (e.g. by income or for geographical areas) to maintain confidentiality, the

¹⁰ The APPSI Trading Fund representative has argued that many new and innovative business opportunities will use this level of detail and any move to make them available at marginal or zero cost would impact on their business model.

information can be made available more widely for public benefit (e.g. to provide some indicators of migration within the UK). The government has agreed the recommendations of the Walport and Thomas report on data sharing¹¹. Given all this, we anticipate wider re-use of aggregate information becoming much more frequent and, indeed, saving money through reducing the need for expensive surveys.

27. Based on all of the above, we believe that there is a strong case for government to review and prioritise its information holdings and, in particular, what constitutes an overall national information infrastructure. This would recognise that government departments would normally have internal requirements for the information they collect but at present there is little by way of a 'whole of government' or 'whole of society' view on what is most important and what should be prioritised for easy access and availability for re-use. Again we recognise some moves in this direction, such as the embryonic developments in creating a planning system for UK official statistics to span the many government bodies which produce them. But we need to have a wider and longer term view of where resources should be put to operate alongside *ad hoc* initiative-driven approaches.

Other government actions

28. We see several other ways in which the cause of PSI re-use could be advanced. These are:

- ***Making PSI held by the whole public sector available for re-use by statute.*** At present the EU Directive translated into UK law does not make this mandatory (though at least one other EU member has taken that step). We **recommend** that some statutory requirement should be implemented;
- ***Sorting out the 'derived data problem'.*** This arises where new information is added to or derived from pre-existing information (e.g. locations of properties or sites of interest digitised from a map even where the particular coordinates are not stored in the data file forming the map). The Intellectual Property Rights issues are presently complex and are a serious constraint on wider use of 'mashing'. We recognise work underway within OPSI to address this issue and stand ready to support it;
- ***Simplifying licensing.*** Much useful work is already underway in this regard which is led by OPSI and involves the Shareholder Executive and Trading Funds. But the problem is deep, pervasive and multi-faceted. A good example where licensing to one community has minimised the value of public expenditure is that the Research Councils UK (RCUK), mainly through ESRC, have developed extensive highly professional software to provide access to public sector information for university academics in the UK. The data in question is already in the public domain including the ONS Population Census data in many varieties and with much value added by the academic Census groups at Manchester, Leeds, Edinburgh and Essex. There is a 'one stop shop' for all this data at the ESRC Data Archive at Essex.

But – despite their pleas - Local Authorities and other users do not have access to this data through these convenient channels even if they already hold a license to the raw Census data. In addition to census information the Edinburgh data service has OS data available to UK academics in terms of some layers of Mastermap; but making this data more widely available is likely to be even more problematic. There are other value added data sets that the UK university community has developed which are available at EDINA and MIMAS such as the LandMap, some fine scale geodemographic data, satellite remote sensing data, the postcode directory etc. Few,

¹¹ <http://www.justice.gov.uk/reviews/docs/data-sharing-review-report.pdf>.

if any of these are generally available thanks to licensing constraints. As another example, even organisations such as the Environment Agency - which already have large holdings of PSI and express commitment to fostering widespread use of 'their' information – have a multiplicity of licensing and other constraints on its re-use which vary from one organisation to another;

- We believe this 'Balkanisation' within parts of the public sector is highly inefficient. Developing common terms for licensing the re-use of information across the wider public sector should help to open up access. It will also reduce the overheads for any body seeking to license such information. Put simply, if an organisation already has a license to access PSI, it should be able to access it via any supplier channel without further complications.

- ***Focusing on all countries in the UK***

The great bulk of discussion about PSI has a strong Whitehall focus. This is unfortunate. We urge both in the short and the longer term that efforts are made to facilitate information being made available for re-use which pertains to Northern Ireland, Scotland, and Wales as well as to England.

- ***Homogenising the rules for Trading Funds.*** Recent developments in regard to Trading Funds have the capacity to reduce the tensions between government and commercial bodies (e.g. through new licensing guidelines being set by OPSI and monitored by OFT). We understand that government will take further steps if these do not succeed but propose that greater homogeneity of approaches in regard to all Trading Funds would be beneficial for those seeking to use PSI from them;

- ***Reducing tensions between business and Trading Funds.*** Beyond that, we see continuing tensions and conflicts between the public service and commercial roles of government departments which impact on the information re-use agenda. One example will suffice to illustrate that. DEFRA and the Meteorological Office have much PSI relating to climate change and the latter has recently summarised this in a widely circulated and well-written document entitled *Warming: climate change – the facts*. This helpfully informs the public about the background to government climate change policy. The relevant section of the Meteorological Office web site also contains the booklet but provides little by way of data. It does however provide access to UK Climate Projections 2009 (UKCP09), a climate analysis tool funded by DEFRA which produces climate projections broken down to a regional level across the UK and shown in probabilistic form - illustrating the potential range of changes and the level of confidence in each prediction. The Meteorological Office states that "Used appropriately and combined with the right information, UKCP09 provides a detailed guide to our future climate";

The reality however is that the web site in general and the expert knowledge of a publicly funded tool which they created under contract is also being used as a marketing tool for Meteorological Office consultancy, as made clear by the Meteorological Office statement that "We can add to the projections in UKCP09 by generating information and solutions tailored to your exact needs. This includes adding more detail to projections by focusing in on specific locations, variables or scenarios, updating projections with the very latest science or taking your specific situation into account." This tension between public and commercial roles of government bodies is at the heart of Trading Funds operations and competition with the private sector. Such issues are certainly not unique to Britain – in Germany, commercial organisations offering a meteorological service have set up their own

network of weather stations after being unable to agree terms with the official meteorological organisation - but are particularly acute here. We believe that such tensions undermine the cause of PSI re-use.

- **PSI and legislation.** We already have recent legislation pertaining to Public Sector Information but, as indicated above, this needs strengthening to be truly effective. We have also observed that the collection, assembly and dissemination of information and statistics by government over many decades are covered by an inconsistent accumulation of legislation that in many cases prevents such information from being made available at all, let alone coherently, consistently and comprehensively. In some cases, information that used to be available is now restricted because of licensing or commercial confidentiality considerations or because new organisations were not mentioned in the original statute. Some mechanisms are now in place to enhance data sharing (e.g. the Statistics and Registration Services Act 2008 provides a Parliamentary mechanism to authorise data sharing). We recognise that up-dating all pre-existing legislation to suit contemporary policies and technological and economic realities would be a mammoth task. But we urge government to ensure that *all forthcoming legislation should make clear not only the responsibilities for collection of relevant information but also a responsibility for its public dissemination*. It is of course mandatory that any confidential personal information would be anonymised before release.

Assembled on behalf of the Advisory Panel on Public Sector Information by David Rhind (APPSI Chairman), drawing on material supplied by APPSI members.