The Foundation for Information Policy Research

Consultation Response on

Making Open Data Real

The Foundation for Information Policy Research (FIPR) is an independent body that studies the interaction between information technology and society. Its goal is to identify technical developments with significant social impact, commission and undertake research into public policy alternatives, and promote public understanding and dialogue between technologists and policy-makers in the UK and Europe.

FIPR welcomes ministers' wish to publish much more government data, both to increase transparency and competition in the public sector, and to let businesses develop new services. Many aspects of this programme are very welcome, from making Ordnance Survey map data more freely available to getting academics to put all their research results online through publishing all government IT contracts.

The big problem is with personal data. Most of us consider our medical records, our tax returns, and our children's school reports to be private, and European law gives us privacy rights in sensitive data that cannot easily be overridden by Parliament.

Some try to square the circle by "anonymising" or "de-identifying" personal data. For years, the NHS has had databases of hospital care episodes with patients' names removed; they are used for everything from academic research to measuring hospital efficiency and surgeons' survival rates. Would it be safe to make such data even more widely available, not just to researchers, healthcare administrators and a few favoured firms, but to any company that wanted it – in effect, to the public?

This would be "courageous" – very unsafe, and wide open to legal challenge.

- It's been known for thirty years that anonymisation is hard. Computer scientists started studying it in the context of the US census after one of the staff bet her boss that she would be able to work out his salary from the data they were planning to publish. The fundamental problem is that if you know more than a handful of things about a person, you can usually identify them.
- There have been detailed studies of "anonymised" medical records by Dr Latanya Sweeney and others, which showed that publicly-released records in the USA could often be re-identified. In the UK, "anonymised" records kept on NHS systems typically still have the patient's postcode and date of birth, which is enough to identify about 98% of patients.
- Things would be better if records had only postcode sector (e.g. CB30 instead of CB30FD) and year of birth, but the Department of Health won't accept this as its statisticians would then be unable to track patients' social deprivation index.

- A more general problem is that good security measures only get implemented if the engineers who build and maintain the system have a real incentive to do so. But medical databases are operated and paid for by researchers and administrators who really want to know things like social deprivation index; their incentive will always be to ask "what's the minimum loss of data we can get away with?" It will always be the least knowledgeable, or least scrupulous, contractor who will give the most pleasing answers to such questions.
- A further problem is that as we move to a world of "big data" where the sheer volumes of data are much greater than a single machine can hold, things become even more difficult technically. Engineers will want, if possible, to work with the original data rather than constantly make redacted versions of it.
- Yet another problem is that systems are becoming "social". All sorts of websites now follow the Facebook model and encourage people to link to their friends – gaming sites, photo sites, and hobbies from genealogy to sport. Patterns of friendships create new context with which people can be re-identified.
- Researchers have built better models of anonymisation, such as "differential privacy" which enables a system operator to measure when her system has answered enough queries that it will start leaking data. This teaches that if privacy is paramount, then after a surprisingly short period of operation, the uncertainty required for anonymity is exhausted – so the database must be shut down.
- For many years, policymakers and lawyers have ignored the advice of technical privacy experts; privacy regulators such as the Information Commissioner usually don't employ any. This wilful blindness is no longer sustainable, because of court cases in the USA and Europe and because legal scholars are beginning to explain in lawyer-readable language what engineers have known for a generation. The most notable such article is by Paul Ohm[1].
- In summary, anonymisation can only be an effective privacy mechanism in rather specialised cases[2]. It is mostly a means of "privacy theatre" used to pretend that systems respect privacy when they don't really. It facilitates regulatory arbitrage: companies can set up "pseudonymous" data in Britain or Ireland, export it to the USA, then re-identify it later for use in advertising or law enforcement.
- Anonymisation certainly cannot resolve the tension between a patient's right to privacy and a medical researcher's desire to have copies of the records of all patients suffering from the disease of interest to her. This is a real social tussle, of the kind that we pay politicians and judges to deal with. Neither should let themselves be fooled into thinking that moving the boundary to favour the data industries, at the expense of privacy rights, will be free of cost, or that there's some technological silver bullet to make it so. A research team has already lost 8.63 million patient records that were given to them in lightly de-identified form[3]. If controls are relaxed further, there will be still more losses.

---

[1] "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" 57 UCLA Law Review 1701 (2010)

[2] 'Security Engineering', RJ Anderson, Wiley 2008, chapter 9

[3] 'London Health Programmes loses unencrypted details of more than 8 million people", W Ashford, Computer Weekly 15 June 2011

The Open Data initiative can bring real benefits. However ministers should focus on publishing non-personal data, and avoid getting dragged into old tussles about third-party access to personal data that will lead to public privacy failures – which, like the HMRC debacle, will undermine public trust.

It may be worth restating the underlying philosophical issue. Information has always tended to flow from the weak to the strong and may further widen the gap between them. The information society exacerbates this problem in two ways. First, the rapidly declining cost of data collection, storage, transmission and processing has led to much greater information flows than before. Second, the information goods and services industries have a stronger tendency than most traditional industries to monopoly, oligopoly and dominant-firm markets, because of network effects, low marginal costs and switching costs. For these reasons we have both data protection laws and freedom-of-information laws, which attempt to redress the balance by giving the weak a limited veto over the flow of their information to the powerful, and compelling (in the public sector at least) some information flow in the reverse direction. If the laudable aim of making the state's data more freely available to the citizen is perverted into a regime that rides roughshod over data protection law to make our personal medical, financial and lifestyle information available to the powerful then the public will rightly lose confidence and hopefully the exercise will be stopped by the European Court.

We therefore answer the consultation as follows. On '**Enhanced Right to Data**':

1. **How would we establish a stronger presumption in favour of publishing than that which currently exists?** We suspect there is no magic bullet here, as the worthy goals of the Freedom of Information Act have been repeatedly subverted by officials using its many loopholes, such as by claiming commercial confidentiality for IT contracts and consultancy reports when these contracts and reports should have been public from the time they were put to tender. The new openness is welcome but ministers' appetite for more openness will fade with time. If there is no will to amend the Act then the enforcement mechanisms had better be improved.
2. **Is providing an independent body, such as the Information Commissioner, with enhanced powers and scope the most effective option for safeguarding a right to access and a right to data?** No. The Information Commissioner's Office was designed from the start to be ineffective. It is kept away from trouble in darkest Cheshire, starved of funds and discouraged from hiring technically capable people. Successive Registrars and Commissioners have not seen their role as protecting privacy rights founded in law, but it helping people get acclimatised to ever-greater flows of information from the weak to the powerful. They provide privacy theatre rather than real privacy, and similarly they provide openness theatre rather than real openness. As we made clear in our reply to the Thomas-Walport Review, the only way we can see to break this cycle is to enable people to take privacy complaints, and freedom-of-information complaints, to the courts.

For that to be practical we must change the rules on costs shifting so that the loser no longer pays the winner's costs, at least in cases founded on human rights.

3. **Are existing safeguards to protect personal data and privacy measures adequate to regulate the Open Data agenda?** No. As noted above, people who suffer privacy infringement need effective access to the courts. Furthermore, the correct way to regulate privacy in Open Data is this: when a department proposes to make personal data available in some de-identified form, the proposal should be published in detail for open public review, perhaps with the personal data of 100 government MPs "de-identified" under the scheme – so that we citizens can kick the tyres and point out the flaws before our own data are put at risk.
4. **What might the resource implications of an enhanced right to data be for those bodies within its scope? How do we ensure that any additional burden is proportionate to this aim?** The best way to do this is that when systems are procured or enhanced, the specifications should be made public (like the contracts) and should set out what data should be public and what should be private. The strong default should be that all non-personal data will be public, while personal data should remain confidential unless the de-identification mechanism has passed a public review as suggested above. Requests for data from existing systems will inevitably be dealt with in an ad-hoc way. However the cost thresholds must be significantly higher than at present, because of the mechanics of outsourcing. For example, the HMRC debacle happened as the facilities management contractor would have charged HMRC thousands of pounds to extract a few records for audit purposes; if contractors' minimum charges for programming work are going to be in four figures, then the freedom of information threshold must be higher still.
5. **How shall we ensure that Open Data standards are embedded in new ICT contracts?** This is for departments to do, and for the public to supervise, now that all IT contracts are going to be published.

On '**Setting Open Data Standards**' we would strongly suggest that standards be set later rather than sooner. Eventually standards will no doubt emerge for large swathes of the public sector (such as healthcare) but until the systems are built and working it won't be clear what they should be. What's more, the civil service already suffers from such onerous internal compliance that it takes months to do something the private sector does in weeks; imposing more of it will be unhelpful, and imposing it on private-sector bodies that fall under the Freedom of Information Act (such as universities) would be perverse as universities are much more open anyway.

'**Corporate and Personal Responsibility**' will first be a matter of setting the right defaults, and secondly a matter of setting the right incentives. Non-personal data such as procurement contracts must be public by default while personal data such as medical, school, tax and welfare records must be private by default. Information flows from private to public must use mechanisms that have been subjected to rigorous open public review, as discussed above. As for incentives, it's important to avoid the problem faced by the Caldicott Guardians in the NHS. This post was established by the Health and Social Care Act 2000 following the Caldicott report into confidentiality in the NHS; a

guardian is a clinician (typically a senior nurse) with line responsibility for patient privacy at a healthcare provider such as a hospital or surgery. The problem is that she has almost no influence whatsoever on privacy; the real privacy decisions are taken in theory by ministers but in practice by the developers who code health information systems. This divorce of power from responsibility is in our view ultimately responsible for the fact that the NHS accounts for by far the largest volume of personal data breaches reported to the ICO in the UK – including the massive leak cited in footnote 3 earlier. The person responsible for privacy must the 'data controller' – the person with effective control over those aspects of system design and operation that determine it. The controller might be a Secretary of State, Permanent Secretary, a Director General, or the Chief Executive of a local authority. Finally, in the absence of any effective sanctions against officials who purchase, deploy and operate systems that unlawfully infringe privacy, it would be perverse to have effective sanctions against any minister or official who failed to publish information against whose release there was some objection on privacy grounds.

'**Meaningful Open Data**' is a desirable goal. It is welcome that departments place data online but the links are forever breaking. Sometimes this is deliberate, as embarrassing crime statistics or documents relating to now-abandoned policies are quietly deleted. Sometimes it results from departmental reorganisations, as when dti.gov.uk becomes bis.gov.uk. Other sources of entropy also contribute. Sometimes third parties keep track, as with www.theyworkforyou.com which provides stable access to the record of Parliamentary proceedings despite the best efforts of Hansard. However this is just not good enough.

We urge the Government to establish permanent URLs for published data. There are many ways to do this, and we'd be happy to advise on the detail. Furthermore, all existing URLs should be grandfathered, so that if the next Prime Minster decides to change bis.gov.uk back into dti.gov.uk, then every single object available at bis.gov.uk should remain available at its current address for as long as .uk continues to exist. In short, the public-sector content management system must be designed not just to last until the next Parliament but until the next civilisation. In fact there should be no need for a separate National Archive; the main production servers should contain the archive, from the earliest documents right up to the present.

'**Government sets the example**' is a fine goal and to get there policy had better be driven by demand. Rather than setting out to scan every out-of-copyright book in the British Library and placing it online (a worthy project but unlikely to be funded in the current climate) the rational approach is to scan books that are requested, and where a book is requested in person rather than electronically, doing the scanning once it's returned so as not to prejudice the level of service. Applying this general insight should help open data champions set priorities. It also suggests that to begin with at least data should be hosted on departmental servers rather than centrally. Perhaps we'll have a 'Government Cloud' in due course but it might take them three attempts and fifteen years to get it right, and it's not sensible for departments to wait.

‘**Innovation in open data**’ is one of the objects of the exercise but is not something that government can do itself or cause others to do directly. People who’re good at this head for industry, academia and the NGO sector. The government’s role is at most a supporting and enabling once.

As for the draft public data principles, these are mostly fine except that the proposal to publish all public data online through a single portal at data.gov.uk is misguided. Portals come and go as ministers come and go and indeed as facilities management contractors do. What’s needed is a system is stable permanent URLs on which others can rely. Until governments ‘gets’ this, it will be open to the criticism that despite the honours bestowed on Tim Berners-Lee, it just doesn’t ‘get’ the web.

Professor Ross Anderson FRS FREng
Chair, Foundation for Information Policy Research
October 27th 2011