

# On Privacy and Public Data: A study of data.gov.uk

Andrew C. Simpson\*

**Abstract.** The web-site `data.gov.uk` (the UK’s counterpart to the US’s `data.gov` site) was launched in January 2010. The site proclaims that “transparency is at the heart of this Government” and that “`data.gov.uk` is home to national & local data for free re-use.” As part of an assignment for a masters-level course on *Data Security* at the University of Oxford, 18 part-time post-graduate students were asked to give consideration to the benefits and drawbacks of releasing public data, with particular focus being given to `data.gov.uk`. In this paper we describe the findings of four of these students and show how the issues raised—both in isolation and when taken in combination—may be a cause for concern, both to those responsible for releasing such data and to those to whom the data pertains. The discussion is not intended to be a critique of `data.gov.uk` *per se*; rather, our hope is that this contribution may play a role in the wider debate pertaining to the issues surrounding the release of public data.

## Keywords:

public data; privacy; deanonymization

## 1 Introduction

The web-site `data.gov.uk` was launched in January 2010 and is “home to national & local data for free re-use.” In many ways, `data.gov.uk` can be considered to be the UK’s counterpart of the US’s `data.gov` site. Both `data.gov` and `data.gov.uk` are part of initiatives to, in the words of a US Government Office of Management and Budget memorandum, “implement the principles of transparency, participation and collaboration.”<sup>1</sup> At the time of writing (February 2011), at least nine countries (the United States of America, the United Kingdom, Canada, Australia, New Zealand, Spain, Norway, Estonia, and Greece) have launched sites that are concerned with “giving data back” to their tax-paying public.<sup>2</sup> It should be noted, however, that such initiatives are not without their critics: as an example, “useless junk” is the characterisation of `data.gov` given by Fetini [10].

These sites typically provide access to data either via download in one of a variety of formats (CSV, XML, etc.) or via links to other sites. According to the `data.gov.uk` FAQ,<sup>3</sup>

---

\*University of Oxford, UK, <mailto:Andrew.Simpson@comlab.ox.ac.uk>

<sup>1</sup>See [www.whitehouse.gov/open/documents/open-government-directive](http://www.whitehouse.gov/open/documents/open-government-directive).

<sup>2</sup>See [www.guardian.co.uk/world-government-data](http://www.guardian.co.uk/world-government-data) for an overview of such data sources from across the globe.

<sup>3</sup>See <http://data.gov.uk/faq/#q9>.

“Excluding personal and sensitive information, all information created by public sector bodies is, in principle, available for re-use. In the past, different approaches were adopted by local and regional authorities and individual agencies. The government is now widely encouraging all previously inaccessible public information to be made accessible through this web-site.”

At of the time of writing (February 2011), the site has over 5,400 datasets available for access—“from all central government departments and a number of other public sector bodies and local authorities.” The types and characteristics of data vary dramatically, as one might expect from an open approach: in some cases, raw statistical data (in the form of CSV files) are available; in other cases, reports including summaries of statistics (in the form of PDFs) are available. In most cases of highly sensitive data, the published data is produced at a high level of abstraction so as to protect individuals; in other cases, some of the datasets feature free text.

It is, perhaps, worth comparing and contrasting `data.gov.uk` and its US counterpart, `data.gov`. The latter launched in May 2009—eight months before the UK site; both are concerned with supporting transparent, open government. As well as raw data, `data.gov` makes a dedicated “Geodata Catalog” available. As well the data, `data.gov.uk`, for example, offers a blog, a wiki, an ideas submission page and a variety of fora; metrics on site usage and content are also available. In some ways, the functionality of `data.gov.uk` goes beyond that of `data.gov`: by making data available as *linked data* [4], there is the potential for machine understanding of the data’s semantics to support the combining of datasets via the query language, SPARQL (see, for example, [18]).

There are key differences with respect to privacy. While, for example, the National Statistics Office has produced a “summary of principles” for data access and confidentiality,<sup>4</sup> no reference is made to this or similar protocols or practices on the `data.gov` web-site; as such, it is unclear as to which types of data are subject to relevant codes of practice. (Although it should be acknowledged that a “Transparency and Privacy Review” has recently been announced.<sup>5</sup>) This is in stark contrast with the US position: the privacy policy associated with `data.gov` is stated clearly.<sup>6</sup> Further, there has been a wider debate with respect to the issues than has hitherto been the case in the UK. For example, the Center for Democracy & Policy Post 15.13 notes that, prior to any release of data, it must be ensured that “the data does not contain personally identifiable information, sensitive information, or other information that could be used to link the released data to individuals.”<sup>7</sup>

As part of a take-home assignment for a course on *Data Security* at the University of Oxford, students were asked to consider the debate of [12], in which the two authors consider the pros and cons of `data.gov.uk`—with one of the authors, Shadbolt, being

<sup>4</sup>See <http://www.ons.gov.uk/about-statistics/ns-standard/cop/protocols/data-access-and-confidentiality.pdf>.

<sup>5</sup>See <http://www.cabinetoffice.gov.uk/news/transparency-and-privacy-review-announced>.

<sup>6</sup>See <http://www.data.gov/privacypolicy>.

<sup>7</sup>See <http://www.opensubscriber.com/message/cdt-announcements@cdt.org/12565308.html>.

a key member of the team responsible for the establishment of the site, and the other, Korff, being a Professor of International Law at London Metropolitan University who has concerns with respect to the threats to privacy and confidentiality associated with the release of such data. In this paper, we consider some of the issues identified by four of the students, and, in particular, consider how the four (related) issues might give cause for concern for those responsible for the release of such data, and, potentially, the subjects of the data. Despite the fact that our study is, necessarily, focused on the situation in the United Kingdom, it is hoped that this paper will have a role to play in the wider debate pertaining to the privacy and confidentiality issues surrounding the sharing of public data.

The structure of the remainder of this paper is as follows. In Section 2 we describe the background to our brief study. Then, in Section 3, we consider the findings of four of the students who submitted the aforementioned assignment. Finally, in Section 4, we summarise the contribution of this paper and provide some concluding remarks. In particular, we give some thought to the potential risks associated with the UK government's drive to encourage local councils to release their data.

## 2 Background

### 2.1 The perceived benefits of releasing public data

The potential benefits of resources such as `data.gov.uk` are described by Omitola and colleagues in [17]. As a specific example, five public datasets, including “datasets of Members of Parliament (MPs), Lords, their corresponding constituencies and counties, relevant web-sites, MPs expenses and votes, and statistical records about crime, hospital waiting time, and mortality rates” are linked to good effect. Further examples of potential applications of the kinds of data that are being made available are given at `data.gov.uk/apps/`; these include a nightly feed of jobs advertised at job centres and a service aimed at those with respiratory diseases that texts people in London details of the air quality in their immediate vicinity.<sup>8</sup>

In many ways, of course, it may be argued that sites such as `data.gov` and `data.gov.uk` are a good thing: the argument goes that taxpayers are entitled to know how their money is spent and such data belongs, in a sense, to the public.<sup>9</sup> Indeed, the UK newspaper, the *Guardian*, has been running a “Free Our Data” campaign for several years now, motivated by exactly these concerns. Certainly, the intentions behind such initiatives are, for the most part, honourable.

The fundamental purpose of `data.gov.uk`, then, is to redress information inequality: to allow people to access and use data which has always been publicly available but might previously have been difficult or expensive to access.<sup>10</sup> To this end, the data available

---

<sup>8</sup>Although, in fairness, the number of applications currently available is rather limited.

<sup>9</sup>Of course, the legal validity of this argument is questionable.

<sup>10</sup>Of course, there are broader concerns with respect to “information inequality”—how might those members of the public who have no Web access benefit from this governmental largesse?—but that,

via `data.gov.uk` is presented in a well-structured and friendly web portal, with further initiatives ongoing with a view to improving accessibility.

## 2.2 Rising privacy concerns

Ironically, the timing of this drive to release public data coincides with greater awareness as to the treatment of personal data by government departments and agencies; certainly, the public's conscience with respect to privacy and data security has been pricked by a series of data-related incidents in the public sector. For example, in October 2007, the entire child benefit database was sent (unregistered and unencrypted) from HM Revenue and Customs (HMRC) to the National Audit Office—only for the disks to fail to arrive. Two months later, it was disclosed that an ex-contractor at the Department for Work and Pensions had been in possession of two disks with thousands of benefit claimants' details for more than a year; the disks contained names, addresses, dates of birth and National Insurance numbers. In the following year, the Home Office was at fault as information (stored on an unencrypted memory stick) on almost 130,000 prisoners and dangerous criminals was lost; it was predicted that the safety of criminals and police informants would be compromised. As a final example, in early 2010, it was reported that the private financial details of up to 50,000 people who claim tax credits had been mistakenly sent out in the post by HMRC—as well as their annual tax credit award notice, recipients were sent personal details of other claimants.

Also in recent years, there has been the publication of various articles with respect to the threats of deanonymization and reidentification—with the contributions of [13], [15] and [21] being pertinent examples in this respect. The most famous example in this area is perhaps the case of the AOL research data release mishap, in which anonymized web searches were made available for research purposes—with one individual being identified in [3] through their searches. Two New York times reporters—Michael Barbaro and Tom Zeller—managed to identify Thelma Arnold from Lilburn, Georgia (user 4417749) by linking “landscapers in Lilburn, Ga”, queries featuring people with the surname Arnold, and “homes sold in shadow lake subdivision gwinnett county georgia”. As an indication of the potentially embarrassing and/or revealing nature of some of the data, Mrs. Arnold also submitted “numb fingers”, “60 single men” and “dog that urinates on everything” as searches to AOL's engine.

Taken together, the goals of transparency and privacy—both of which have risen up the public agenda in recent years—can be seen to be in conflict. The tension between utility and confidentiality is captured wonderfully in [11]:

“Arguably, the most urgent need is not for development of more techniques for SDL [statistical disclosure limitation], but for research that provides agencies methods and tools for making sound decisions about SDL. There is a broad consensus that, in principle, releasing data—either literally or through a set of allowable queries (and possibly incomplete responses)—

---

one supposes, is not part of this particular agenda of widening participation.

is a decision problem in which each release is characterized by quantified measures of risk and utility from which a principled choice can be made in several ways. But, inability to implement this paradigm remains nearly total.”

And, of course, as evidenced by [1], these issues span a wide range of disciplines:

“Statisticians, particularly those working within national statistical offices, have developed the field of statistical disclosure limitation. Computer scientists contribute work in privacy-preserving data-mining and cryptographic analyses of privacy. Lawyers and social scientists study the role of government and regulation in the creation and protection of individual and business privacy. Health researchers struggle with the trade-off between a patient’s privacy and the contribution to science that access to integrated medical records might allow. Survey designers in all fields of human endeavor wrestle with methods of enticing survey cooperation under a variety of ethical and privacy guarantees.

“Gargantuan online services gather petabytes of data on search queries, online purchases, e-mail exchanges, and other social network interactions while pushing their computer scientists to exploit the corporate asset these data represent without damaging the companies’ ability to do future business by breaching the confidence of their client/users. And many, many data users from all of the fields listed above perform analyses that are conditioned on the privacy and confidentiality protections imposed on their work without all the means to assess the consequences of those measures on the inferences they have made.”

Other references in this regard include [16] and [7]. As a further example, [19] considers the balance between confidentiality and functionality in the context of medical research. Taken to the extreme, [6] states: “it may be argued that elimination of disclosure is possible only by elimination of statistics.”

It is this tension—between functionality and privacy—that underpins both the contribution of this paper and the wider debate with respect to the benefits and drawbacks of releasing public data.

## 2.3 The debate

In [12], two protagonists, Nigel Shadbolt and Douwe Korff, consider the pros and cons of `data.gov.uk`. To a first approximation, the debate centres around Shadbolt’s argument that releasing public data is undoubtedly a good thing and can be the source of innovation (the potential to build applications to “tell you where to cycle to avoid the accident black spots” or to “locate your nearest NHS dentist”), while Korff gives consideration to the potential breaches of privacy that may arise. Arguments put forward for the release of data include the desire to “redress information inequality” with a view

to giving data back to the public as “data offers remarkable opportunities to empower citizens.”

The volumes and types of government data that is “personal”, as opposed to “non-personal” is disputed: Korff argues that almost all data is personal to some extent as it is commonly derived from the actions of individuals who can be identified, while Shadbolt argues that aggregate, anonymized data is “non-personal public data” and belongs to those upon which it is collected.

Suggested controls to try and ensure that breaches of privacy<sup>11</sup> are limited amount to legal and ethical restrictions: “we need good law and regulation, social conventions and behavioural norms that respect personal information.” However, of course, once data is released, it is, to coin a phrase, “in the wild”; there is no avenue of recall. Indeed, as Shadbolt and his co-author O’Hara point out in their textbook, *The Spy in the Coffee Machine*: “digital information lasts a long time, effectively forever if it is periodically copied, backed up and stored using up-to-date forms” [14]. Going further: “searching through digital information is fast; discovering a tiny number of references to a person in a large database, virtually impossible to spot with the human eye, is a simple matter with a computer. Information that is harmless on its own can be placed in significant new contexts.”

For the most part, [data.gov.uk](http://data.gov.uk) publishes only datasets containing aggregate statistics about people.<sup>12</sup> However, even when we limit ourselves to statistical data, it is well understood that privacy attacks (via, for example, *trackers* [8]) are possible; when considering microdata, concerns pertaining to deanonymization (see, for example, [15]) come into play.

Some of the concerns surrounding “joining the dots” between apparently disparate data sources are as much to do with *mis*-identification as they are to do with re-identification. An example of a false positive is that of Michael Hicks: an 8-year old cub scout from New Jersey who has to face extra security checks at airports due to the fact that he has the same name as a terrorist [5]. With respect to potential dangers of misidentification in the UK, there are precedents—with vandals confusing the terms *paediatrician* and *paedophile*, leading to a doctor being forced to move home, being a particularly notable example [2].

## 2.4 The Data Security assignment

The assignment question (one of two) was given as follows:

“First, consider the discussion between Korff and Shadbolt on the pros and cons of [www.data.gov.uk](http://www.data.gov.uk). Second, download some datasets from [www.data.gov.uk](http://www.data.gov.uk) to get a feel for the kind of data that is available and the potential privacy issues involved.

---

<sup>11</sup>There is, though, some debate between the authors as to what the term “privacy” actually means.

<sup>12</sup>Although, as we shall see, there are examples of datasets that don’t meet this description.

“Describe where you stand on the discussion. You should feel free to make use of any resources to support your arguments; a strong answer should both leverage relevant literature and give consideration to some of the datasets available from the site.

“Remember that a scientific argument needs to be backed up with evidence (don’t simply give opinions!), and that you should describe the strengths and weaknesses of each position before presenting a reasoned conclusion.

“It would be surprising if an answer were to be longer than 5 pages.”

The one-week intensive course (aimed at professional software and security engineers, studying for a part-time MSc in one of Software Engineering or Software and Systems Security) was attended by 18 students in June 2010. 15 of the 18 students submitted their assignment six weeks later.

In the following section, we consider the findings of four of those students.

### 3 The results

Of the 15 submissions, nine responses were of the form “I see no problem here—sharing public data is a good thing,” while two were of the form “it feels like there’s a problem here—but I can’t find any evidence.” Four of the submissions gave evidence of concerns. Interestingly, each of these four candidates identified different problems—some more significant than others; yet taken together, there is the potential for significant mischief—especially if the proposed interfaces to link data from `data.gov.uk` and other sources were to be leveraged effectively.

Our first issue pertains to limitations on statistical data—which forms the core of the provision of `data.gov.uk`.

#### 3.1 Limits on statistical data

Our first student suggests that “one might expect that `data.gov.uk` would employ some very robust criteria in selecting and/or modifying its datasets to prevent re-identification, and dedicate considerable web-space to reassuring its critics of them.” Following some investigation, however, it is discovered that this is not always the case: “for example, we know that only one individual in the North-Tyneside PCT attended an appointment after 10 days for a genito-urinary tract infection in May 2010.”

Of course, if only statistical data were to be made available—with appropriate *a priori* protection via sampling—then the potential for privacy breaches would be limited. However, while this is mostly the case, it is certainly not true in all cases. One might argue that the apparent lack of guiding principles with respect to data release that was alluded to in Section 1 might be a cause for concern in this respect.

### 3.2 Poorly anonymized datasets

Our second student discovered datasets that included free text data that had been entered by individuals: in some cases, the data contained identifying (or at least partially identifying) information. One example given was the *General Practitioner Comments and Responses* dataset. In the dataset, there are examples of both patients and GPs identifying each other (as well as other parties) directly by name; also included is supplementary information that may be used to identify individuals. Examples given by the student include the following:

- “Despite patient not including her name in the message, the GP’s response directly refers to her by her first name.”
- “A GP’s response identifies a patient’s surname, who is claiming that the only good thing about the surgery is its ‘proximity to [his] house’.”
- “A patient is identified by name by both himself and the GP’s response”; “additionally the patient provides information indicating that he is approximately 40 years or older.”

The student subsequently makes the argument that linking these fragments of data with other datasets would not be a great challenge. As an example, the *SaferMK* website<sup>13</sup> represents organisations working with the community to try and help reduce crime, anti-social behaviour, etc. In particular, the *SaferMK* Community Safety Mapping application “provides crime and anti-social behaviour data for every estate, town or village in the Milton Keynes borough.” It is possible to find distinct elements of the dataset that the *SaferMK* app consumes (a single hate crime in Loughton in June 2010; a single sexual offence in Woughton South in July 2010).

It follows that there is the potential to link statistical data with poorly anonymized microdata from other sources. An example along these exact lines was provided by a third student.

### 3.3 Linking statistical and personal information

Our third investigative student discovered that there is personal information available from [data.gov.uk](http://data.gov.uk) that does identify individuals: names and addresses of councilors, and names, posts and salaries of senior crown employees, for example.

Some government departments provide their own statistical web-sites and related services, with one example being the Ministry of Defence (MoD). Via the MoD’s site, it is possible to create custom queries. The student ran a query asking for the number of personnel in each service by rank, resulting in the following table:

---

<sup>13</sup>See <http://www.milton-keynes.gov.uk/safermk/>.

| <b>Rank</b> | <b>Total</b> | <b>Royal Navy</b> | <b>Army</b> | <b>Royal Air Force</b> |
|-------------|--------------|-------------------|-------------|------------------------|
| OF-10       | –            | –                 | –           | –                      |
| OF-9        | 10           | -                 | -           | -                      |
| OF-8        | 30           | 10                | 10          | 10                     |
| OF-7        | 100          | 30                | 40          | 30                     |
| ⋮           | ⋮            | ⋮                 | ⋮           | ⋮                      |

(We concentrate here on the figures for senior ranks.)

In the explanatory notes, it can be found that:

- ‘–’ denotes 0 or rounded to 0, and
- the figures have been rounded to the nearest 10.

The 10 officers of rank OF-9 correspond to General (Army), Admiral (Royal Navy) and Air Chief Marshal (Royal Air Force), with the explanatory notes indicating that there are less than five individuals of this rank per service.

From data published by the Cabinet Office, the student discovered a file with the pay details of 345 Senior Civil Servants, Military Officers and Government Advisors. The file listed 10 Military Officers, only 9 of whom were of OF-9 rank:

|                       |                   |                     |
|-----------------------|-------------------|---------------------|
| Sir Jock Stirrup      | Air Chief Marshal | £240,000 – £244,999 |
| Sir Mark Stanhope     | Admiral           | £175,000 – £179,999 |
| Sir Kevin O’Donoghue  | General           | £175,000 – £179,999 |
| Sir David Richards    | General           | £165,000 – £169,999 |
| Sir Stephen Dalton    | Air Chief Marshal | £165,000 – £169,999 |
| Sir Nicholas Houghton | General           | £165,000 – £169,999 |
| Sir John McColl       | General           | £165,000 – £169,999 |
| Sir Trevor Soar       | Admiral           | £160,000 – £164,999 |
| Sir Peter Wall        | General           | £160,000 – £164,999 |

Thus the numbers of offices per service that met the criteria could be determined: for the Army it was 5; for the Navy it was 2; and for the Air Force it was 2.

Through further research, the student found the 2008 pay review by the Review Body on Senior Salaries. Page 9 of this publication details the pay bands for OF-9, with an additional table for the pay bands for the Chief of the Defence Staff. As the student concluded:

“It is therefore possible to identify the seniority of the officers. Now that names and posts it would be possible to expand research of these individuals to web-sites and sources outside of the statistical and data web-sites looked at.”

The individuals whose details have been published all hold senior posts—data was published with the intent of providing a comparison between remuneration in the private and public sectors. The lack of consideration of “other” data in the public domain prior to publication of anonymized data has led to the identification of an individual. Now, of course, the data pertaining to salaries was already in the open; the fundamental point here is that by using data that was already in the public domain, efforts at disguising the numbers released by the MoD are futile.

Our final example gives consideration to live data feeds—and how they may be correlated with other data or previously concluded insights.

### 3.4 On energy use and national security

The `data.gov.uk` site releases data pertaining to energy consumption data (both live and historic) for some government departments, one of which is the MoD. Views of the data are offered in a variety of time intervals (half-hourly, daily, weekly, monthly, quarterly and annual).

In trying to determine whether there was any correlation between this data and (potentially sensitive) government activities and military movements, the student discovered the following:

“For the latest day data is available (Friday 6 Aug 2010), energy peaked at 0730–0800 (2,342 kWh) and began to tail off significantly at 1630–1700 (2,168 kWh to 1,762kWh), as might be expected with a workday. This data over time could have the potential to identify primary time frames of employee movement and permit the inference of the staffing level of the building at a given point in time. It would not be difficult to further correlate this information to get more precise data (e.g. the publicly reported manning numbers for the MoD).”

Taking this argument forward:

“Looking at the historic data available since 8 July 2010 in half-hour intervals ... some identifiable aspects include:

- (i) The five highest dates of raw power consumption are 12 July, 19 July, 8 July, 9 July, and 13 July, respectively.
- (ii) One possible explanation for any surge in consumption could be an increased need in cooling requirements. The five highest temperature days for London during this timeframe were 9 July, 10 July, 19 July, 20 July, 8 July, respectively. As 10 July is a weekend day (where the building could be assumed to house less staff, and require less power) we can assume a power increase due to cooling requirements factor (but perhaps not fully explain) the increase for 8 July, 9 July, 19 July; how-

ever 12–13 July (high temperatures of 22 and 21, respectively) would seem to be outliers to this trend.

- (iii) Interesting surges of consumption are visible for the time period 20:00–06:00 for the dates 11–12 Jul and 18–19 Jul (weekday, non-workday hours). In both cases, a significant increase of electricity can be seen (average of 352.47 kWh difference in these two totals over the third highest total for these dates and timeframes, compared to an average a difference of 76.44 kWh between the third and fourth highest)."

Going further, web searches by the student revealed the following:

- (a) "Lead stories by the *Guardian* newspaper involving recent events involving the MoD include 8 July, 14 July, 19 July, 20 July, and 28 July."
- (b) "Beginning dates of operations listed on the MoD web-site for this time span are 30 Jul and 19 Jul."

While no definitive conclusions can be made from this limited data, some clear correspondences in dates were noted—with 8–9, 12–13 and 18–19 Jul being examples. As the student readily acknowledges, there are precedents for correlating such data, with an example being the episode described in [9]. The article describes interviews with pizza delivery personnel who claimed to be able to predict the announcement of major military operations based upon pizza delivery orders to places such as the Pentagon, the White House, and the CIA. The point is made that the energy consumption data is not limited to the Ministry of Defence, but includes other departments (including the Home Office and the Ministry of Justice)—meaning that analysis of energy consumption across departments could give rise to the potential for unwanted information flow, perhaps indicating a national crisis of some sort.

### 3.5 Summary

In [12], Shadbolt claims that all of the information that is now available through **data.gov.uk** has always been available previously to the rich and powerful—and that initiatives such as this serve only to, in some sense, "level the playing field". While the democratic argument is, on the face of it, a convincing one, the fact that pseudo-personal data can now be accessed with ease—from anywhere, with no real accountability—is significant: the inferences with respect to national security developed on the basis of energy consumption data were established by a student thousands of miles away from the UK. Further, there are no restrictions in place to ensure that only those from the UK (in the case of **data.gov.uk**) or the US (in the case of **data.gov**) can access data—which rather weakens the "giving the data back to the tax-payer" argument.

In all fairness, we should acknowledge the distinction between *disclosure risk* and *harm*: if the data is already in the public domain then one might conclude that there is no disclosure risk—but there is the potential for harm. For example, obtaining the

salaries of high-ranking officers may cause harm, but there is no risk of disclosure. On the hand, the example of Section 3.3 demonstrates genuine disclosure risk—as the student was able to attack the table published by the MoD to decode the rounding.

As we have seen, there are potential dangers associated with this “rush to publish”—with these dangers being associated primarily with the linking of disparate datasets. Consider, for example, the web-site [www.192.com](http://www.192.com), which gathers data pertaining to UK citizens from local councils. Within seconds, one may find details of: the current author’s address and telephone number, the duration that she has lived at his current address, the price that she paid for her house, names of current and former co-occupiers, birth details, marriage details, etc.—all packaged together in one convenient, accessible location. It might be argued that any distaste for the existence of such sites is tangential to the issues surrounding [data.gov.uk](http://data.gov.uk)—which would certainly be true if only non-personal data was released, or if such data was perfectly anonymized, or if accepted approaches to statistical disclosure were followed consistently, or if one were unable to make inferences from released data. But, as we have seen, it is unfortunately the case that each of these problems exists.

## 4 Conclusions

The potential benefits of a system such as [data.gov.uk](http://data.gov.uk) are significant: both in terms of giving the impression of transparent and accountable government and with respect to supporting innovation. Arguments about whether these benefits are actually being realised (especially for the “ordinary taxpayer”) are debatable. Further, these potential benefits are balanced out by the potential for abuse and misuse of data. In the space of a few weeks, a group of (admittedly highly competent) students detected potential problems with data available from [data.gov.uk](http://data.gov.uk).

Returning to the discussion of [12], the claim is that it is a project to “make non-personal public data public,” with examples of non-personal data being “information about the weather, the state of our roads, the physical and administrative geography of the country, environmental emissions levels, what our taxes are spent on and so on”; as we have seen, this is far from the whole story.

There have, to date, been (to the best of the author’s knowledge) no significant privacy attacks on [data.gov.uk](http://data.gov.uk). Nevertheless, the claim in [12] that we “need good law and regulation, social conventions and behavioural norms that respect personal information” borders on the naïve: many individuals are perfectly prepared to step outside the boundaries prescribed by “behavioural norms” on a regular basis. (The argument that social norms, law and regulation will eventually catch up with technology is also made in [20] (albeit in a slightly different context).) Also, of course, by offering the data for download, any effective notions of accountability and auditing are lost.

Furthermore, as the datasets are available over the Internet, any “social norms” associated with the UK (or the US in the case of [data.gov](http://data.gov)) will not apply overseas. Indeed, as our fourth student (of Section 3.4) discovered,

“it is clear that the data isn’t just of interest to those who paid for it (or upon whom it is collected); examining the number of hits for the period January 2010–June 2010 ... shows that the top 10 non-US countries (the set for which foreign data is reported) account for anywhere from 10% to a whopping 33% of the total number of hits in each month over this timeframe.”

Going further:

“An examination of the top downloaded datasets (for the last 30 days and all-time) reveals datasets as diverse as US Overseas Loans and Grants (detailing US foreign assistance), Travel Warnings (detailing assessments of country conditions and ability to assist citizens due to embassy closure or staff drawdown), Active Mines and Mineral Plants in the US, and EPA Geospatial Data Download: Facility and Site Information.”

In Section 1 we acknowledged that our discussion was limited to the situation within the UK, and it was our desire that our account should have relevance to other territories. There are, though, a set of circumstances which are unique to the UK. For example, the move towards an era of “transparent government” is driven partially by pragmatic, political concerns: it is in part a reaction to the recent scandal pertaining to the expenses system for Members of Parliament (MPs)—which a significant number of MPs were found to be exploiting (some to a degree that has subsequently been deemed to be criminal).

Recently, there has been a drive to release even more data in the UK, driven by the newly incumbent Conservation-Liberal Democrat coalition government. For example, the Communities Secretary, Eric Pickles, recently urged in an open letter “all councils to publish details of all spending over £500 in full and online as part of wider action to bring about a revolution in town hall openness and accountability,”<sup>14</sup> and councils “being encouraged to throw open their files and publish, alongside spending data, information on salaries, job titles, allowances and expenses, minutes of meetings and more.” Examples of data to be published under these directives include: local government salaries; councillor allowances and expenses; council minutes and papers; licensing applications and decisions; planning applications and decisions; and food hygiene reports for food outlets. The relatively gung-ho attitude to this initiative is captured by the phrase “I don’t expect everyone to get it right first time, but I do expect everyone to do it.”

The release of data by central government departments and agencies is one thing: there are significant available resources to help protect privacy—in terms of, for example, domain knowledge, statistical analysis, and technology provision; the release of data at the local level—where such resources will not be as readily available—is potentially more harmful. Of course, determining what may or may not be appropriate to be released—and what may lead to harm—is context-sensitive: the potential to build up a threat

---

<sup>14</sup>See, for example, [www.guardian.co.uk/news/datablog/2010/sep/10/local-council-spending-over-500-list](http://www.guardian.co.uk/news/datablog/2010/sep/10/local-council-spending-over-500-list).

model—in terms of what “attackers” might already know, or more to the point, what may be available to them to compromise privacy—is an extremely difficult task. As such, a cautious approach—favouring the privacy side of the functionality vs. privacy tension of Section 2—would appear to be appropriate. Certainly, departments and organisations releasing complementary data independently with limited consideration as to what else is “out there” would appear to be a recipe for disaster. Unfortunately, however, such a conservative approach would seem to be inconsistent with “I don’t expect everyone to get it right first time.”

### Acknowledgments

I am indebted to the four students—Christopher Blake, Jonathan Crew, Chad Heitzenrater and George Svarovsky—whose thoughtful submissions to their Data Security assignment prompted the writing of this paper. I am also grateful to both Jonathan Crew and Mark Slaymaker for their insightful comments.

### References

- [1] Abowd, J. M., Nissim, K., and Skinner, C. J. (2009). First issue editorial. *Journal of Privacy and Confidentiality*, 1(1): Article 1.
- [2] Allison, R. (2000). Doctor driven out of home by vigilantes. *The Guardian*, August 30.
- [3] Barbaro, M. and Zeller Jr., T. (2006). A face is exposed for AOL searcher no. 4417749. *New York Times*, August 9.
- [4] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data—the story so far. *International Journal on Semantic Web and Information Systems*, 5(3): 1–22.
- [5] Bone, J. (2010). Meet Michael Hicks—the boy who has been a terror suspect since he was 2. *The Times*, January 16.
- [6] Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15: 429–444.
- [7] Damiani, E., De Capitani di Vimercati, S., Jajodia, S., Paraboschi, S., and Samarati, P. (2003). Balancing confidentiality and efficiency in untrusted relational DBMSs. In *Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS '03)*, 93–102.
- [8] Denning, D. E. and Denning, P. J. (1979). Data security. *ACM Computing Surveys*, 11(3): 227–249.
- [9] Ellis, D. and Gray, P. (1990). And bomb the anchovies. *Time*, August 13.
- [10] Fetini, A. (2009). Fulfilling a campaign promise: Better access to useless junk. *Time*, May 28.

- [11] Kinney, S. K., Karr, A. F., and Gonzalez Jr., J. F. (2009). Data confidentiality: The next five years summary and guide to papers. *Journal of Privacy and Confidentiality*, 1(2): Article 1.
- [12] Korff, D. and Shadbolt, N. (2010). Public information: Cause for celebration or concern? *Public and Science*, 10–11.
- [13] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 111–125.
- [14] O’Hara, K. and Shadbolt, N. (2008). *The spy in the coffee machine: the end of privacy as we know it*. Oneworld.
- [15] Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. University of Colorado Law Legal Studies Research Paper 09–12, University of Colorado Law School.
- [16] Olivier, M. S. (2002). Database privacy: balancing confidentiality, integrity and availability. *ACM SIGKDD Explorations Newsletter*, 4(2): 20–27.
- [17] Omitola, T., Koumenides, C. L., Popov, I. O., Yang, Y., Salvadores, M., Szomszor, M., Berners-Lee, T., Gibbins, N., Hall, W., Schraefel, M. C., and Shadbolt, N. (2010). Put in your postcode, out comes the data: A case study. In *The Semantic Web: Research and Applications*, vol. 6088 of LNCS. Springer-Verlag. 318–332.
- [18] Quilitz, B. and Leser, U. (2008). Querying distributed RDF data sources with SPARQL. In *The Semantic Web: Research and Applications*, vol. 5021 of LNCS. Springer-Verlag. 524–538.
- [19] Wartenberg, D. and Thompson, W. D. (2010). Privacy versus public health: The impact of current confidentiality rules. *American Journal of Public Health*, 100(3): 407–412.
- [20] Weitzner, D. J., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., and Sussman, G. J. (2008). Information accountability. *Communications of the ACM*, 51(6): 82–87.
- [21] Wondracek, G., Holz, T., Kirda, E., and Kruegel, C. (2010). A practical attack to de-anonymize social network users. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, 223–238.

