

# Public information: cause for celebration or concern?

Douwe Korff and Nigel Shadbolt disagree

January saw the launch of a government website, [www.data.gov.uk](http://www.data.gov.uk), hosting 2,500 public data sets on subjects as diverse as football banning orders by club, the number of visits to museums, the house price index, and bus punctuality statistics. Is the release of this and other government data a cause for celebration or concern? Douwe Korff and Nigel Shadbolt argue it out.

## Dear Nigel,

You argue that most government-held data should be 'freed' and easily-linkable. You cite, as an example, the amount and type of traffic on our roads, where the accidents are, and how much is spent on areas where these accidents occur. However, you should consider the effects of such a policy on individuals to whom the data relate, or to whom the data would be applied.

Many data held by government relate to identified individuals. Just think of the NHS and its new, contentious, centralised systems. Many of the 'non-personal' data are derived from individual incidents, relating to real, identifiable people.

It will be impossible – and I mean technically and mathematically impossible, not just difficult – to free all non-personal data without effectively also releasing (or allowing the re-construction) of vast amounts of identifiable, personal – indeed, often sensitive data.<sup>1</sup>

Non-personal data are also increasingly used to create profiles – companies use them to screen job applicants or identify 'probable' fraudsters; governments screen immigrants and identify 'probable' terrorists (or paedophiles, or youngsters that don't eat enough green vegetables). The forthcoming census will provide the greatest ever resource for such policies. How would you prevent the use of 'non-personal' data in such ways?

Yours, Douwe

<sup>1</sup> See, for example, P Ohm (2009), *Broken promises of privacy: Responding to the surprising failure of anonymisation*. Working paper number 09-12, Legal studies research paper series, University of Colorado Law School

## Dear Douwe,

Tim Berners-Lee and I are leading a project to help make non-personal public data public. We believe this will create social and economic value, make government more accountable and help improve our public services. A huge amount of data held by Government really is non-personal – information about the weather, the state of our roads, the physical and administrative geography of the country, environmental emission levels, what our taxes are spent on and so on for thousands of datasets.

The release of this data should not be contingent on the notion that it might in some tenuous way relate to an individual. Certainly accidents relate to people. But anyone can sift the local newspapers and quickly come up with lists of actual people, injured or killed in traffic accidents. Technology makes this easier but it has always been possible.

I have argued long and hard about the importance of privacy in our technological age – see *The Spy in the Coffee Machine*.<sup>2</sup> In the past our privacy was ensured by a kind of practical obscurity. It was too hard to find and pull the information together. We need good law and regulation, social conventions and behavioural norms that respect personal information.

Yours, Nigel

<sup>2</sup> Kieron O'Hara and Nigel Shadbolt (2008), *The Spy in the Coffee Machine: The End of Privacy as We Know it*. Oneworld Publications

## Dear Nigel,

*The Spy in the Coffee Machine* takes a very American view of privacy as just 'the right to be left alone'. By contrast, European data protection law focuses on the use of data to exercise control over individuals.

In spite of what you say, much 'non-personal' information does stem from data on individuals, and if very large datasets are 'freed', it becomes impossible to keep the data anonymous (see Ohm<sup>1</sup>). This is incomparable to a trawl through paper clippings. Encryption (to which *The Spy* devotes much space) is useless against this.

'Non-personal' data can be highly revealing of individual (or household) behaviour.<sup>3</sup> Proposals for 'smart' energy meters, still pursued here, were defeated in the Netherlands because of this.

Population-scale databases are increasingly interconnected and mined to create profiles linked to bad things: obesity, teenage pregnancies, 'extremism'. Increasingly, 'the computer' takes decisions that significantly affect individuals, on the basis of dynamically-created algorithms extracted from those data, that even the officials or staff implementing the decisions do not understand, and that data subjects are unable to challenge.

If you free all datasets, this will lead to more profiling, with chilling effects on individual freedom and a reduction in democratic accountability.

Yours, Douwe

<sup>3</sup> See Elias L Quinn, Privacy and the new energy infrastructure, <http://ssrn.com/abstract=1370731>.

### Dear Douwe,

First of all we are not publishing 'all data sets'. To be clear, non-personal public data is the scope of data.gov.uk, whose launch was greeted with huge enthusiasm.<sup>4</sup>

Data about your neighbourhood statistics, accident and crime rates, educational attainment, road usage (all real examples we have unleashed on the website) begin to redress information inequality. If you were rich enough, a big enough business, powerful enough and had all the time in the world you might have been able to get hold of some of it before. We are giving this data back to the public who paid for its collection in the first place.

People take this data and do remarkable things with it – build applications to tell you where to cycle to avoid the accident black spots, or locate your nearest NHS dentist. Applications show how far you can live from your place of work if you have a fixed journey time, and what it would cost to live there.

Aggregate (non individual) statistics about mortality, obesity and pregnancy drive health and social policy. I believe the public would like to know what those facts are so that we can hold our governments accountable.

### Yours, Nigel

4 [www.guardian.co.uk/technology/datablog/2010/jan/21/government-free-data-website-launch](http://www.guardian.co.uk/technology/datablog/2010/jan/21/government-free-data-website-launch)

### Dear Nigel,

We agree that releasing government information is a good thing, and that data protection is important. But claiming that the data on data.gov.uk are 'non-personal' is not good enough. If your data are derived from real-world datasets of individual incidents (such as traffic accidents or burglaries), then it becomes almost impossible to prevent the full re-identification of the data by someone with access to even a small number of identifiers (like age or house type).

Some measures counter very simple 'matching', such as trivially-de-identified hospital and electoral records. However, near-re-identification becomes possible when you can match large amounts of even seriously-anonymised data across many large datasets. Narayanan and Shmatikov have worked out the relevant algorithms.<sup>5</sup> And when companies and state agencies make quasi-matches, they will consider them as real, and will treat the target as a likely customer, or fraudster or terrorist. They will also use your datasets to create or improve the 'profiles' they already have. Are you happy with that?

You say that you do not publish all datasets. However, the US website that inspired you offers complete 'raw' data, see: [www.data.gov/catalog/raw](http://www.data.gov/catalog/raw). If you do not want to do that, where do you draw the lines?

### Yours, Douwe

5 Arvind Narayanan and Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, University of Texas, 2008, available from: [http://www.cs.utexas.edu/~shmat/shmat\\_oak08nefflix.pdf](http://www.cs.utexas.edu/~shmat/shmat_oak08nefflix.pdf)

### Dear Douwe,

We agree that releasing government information is a good thing. Excellent! We agree upholding privacy is a good thing. Great! Your concern is over 'information triangulation' – the more information you have, the more it is possible to associate this information with individuals.

I do not contest the power of the methods you describe. But this is not about technology or computational algorithms. I am with Wiezner<sup>6</sup> – what we want is accountability. If data is used in this way then we need to hold people, companies and agencies accountable.

Public opinion can do this. Witness how quickly Facebook changed its policy about not deleting the data of those who left the site. Regulation and law can also hold one to account.

We don't allow companies to look inside the packets of data we ship across the Internet. Our common law is what we commonly choose to uphold.

Public data for the public is a public good. Shroud waving about the dangers of making it available is to miss the point. We certainly do need to determine the rules and limits in an information intensive world, but I believe releasing non-personal public data offers remarkable opportunities to empower citizens.<sup>7</sup>

### Yours, Nigel

6 D J Weitzner et al Information Accountability June 2008/Vol. 51, No. 6 Communications of the ACM

7 <http://www.guardian.co.uk/commentisfree/2010/jan/23/government-information-creative-commons-internet>



**Douwe Korff** is Professor of International Law at London Metropolitan University [d.korff@londonmet.ac.uk](mailto:d.korff@londonmet.ac.uk)

**Nigel Shadbolt** is Professor of Artificial Intelligence and Deputy Head (Research) of the School of Electronics and Computer Science at the University of Southampton [nrs@ecs.soton.ac.uk](mailto:nrs@ecs.soton.ac.uk)