When the chips come out: is our public interest research infrastructure fit for the future?

Jen Persson
jen@defenddigitalme.com

Jen Persson has been a lay member of the Administrative Data Research Network (ADRN) Approvals Panel since April 2015. This paper expresses her personal opinion and work.
She is coordinator of the children's civil liberties campaign group, defenddigitalme.

Abstract

Opportunities for research using more population-wide datasets are within sight in new UK and EU legal frameworks, but inconsistent policy and practices continue to jeopardise data access and public benefit.

Failure to use data in ways the public expect, to safeguard data adequately, and to engage with concerns over consent and confidentiality beset care.data plans and led to a breakdown in public trust in 2014. These effects have disrupted public interest research since.

To avert similar contagion in other areas of administrative data, repeating mistakes must be avoided.

Policy decisions are incrementally expanding children's data collection and use, linking health with education data, and wider data sets. Different pathways provide access to data to a wide variety of third parties.

Exploring public awareness of confidential pupil data in the Department for Education's (DfE) 20m strong National Pupil Database (NPD) we consider research infrastructure in England — data access routes and users, and its foundation on public trust and legislation.

We gathered qualitative responses from 75 schools, 100 education practitioners, 100 parents of children aged 2-19, and from 25 students aged under thirty-five. We found familiarity with school census collection, but none with where data goes once it leaves local systems. People were surprised by the release of sensitive identifiable data to third parties, and that journalists, charities and commercial users received data since 2012.

Change is needed from policy makers and practitioners making our infrastructure fit for data, and a smart future.

Keywords – public engagement; privacy; civil society; ethics; data science

Introduction

In December 2013 a geneticist told the House of Commons Education Select Committee, when the chips come out that can identify people's DNA differences, it's going to really change things fast. (Underachievement in Education, House of Commons Education Select Committee Report, 2014) Plans abound for children's databases, (CHIS) and the ISCG approved part-of-care.data plans, the 2015 mandated collection of the Maternity and Child Dataset by NHS Digital (formerly the HSCIC, Health and Social Care Information Centre).

The public expect their data to be safe and used transparently with 'no surprises' (Wellcome, 2015), in order to secure a social license for research (Carter et al., 2015). Yet population-wide datasets continue to grow unseen, the scope of uses and users expanding over time.

There are high hopes for the expansion of public data access through new legislation in the Digital Economy Bill 2016, and significant investment in a UK wide safe research data infrastructure. However opportunities are at risk as long as public trust is underplayed or undermined, and data are used in ways the public do not expect. The health data secondary uses opt out promised in 2014, and enacted through the HSCIC in March 2016, seems, for example, uncertain once again. (Caldicott consultation, New data Security Standards and Opt-out Models for Health and Social Care, August 2016)

More administrative data use may become increasingly compromised, if it is felt 'Big Data has rendered obsolete the current approach to protecting privacy and civil liberties.' (Mundie, 2014).

The approach to handling data in safe settings contrasts with the release of identifiable data into-the-wild. Consistent safe policies — standards and oversight how public data not only 'can be' used, but 'should be' used, accommodating consensual data subject rights — are needed across public data to future-proof public trust.

1 The National Pupil Database

The National Pupil Database (NPD) is one of the richest education datasets in the world and holds a wide range of information, extracted since 2000 from pupils aged 2-19 at the time of collection. It includes a number of different data collections from schools, Local Authorities and awarding bodies, processed by the DfE's Education Data Division (NPD Guide, 2015, p5).

1.1. Database size, unpublished numbers

We obtained the size of the database through Freedom-of-Information. 'The total number of Unique Pupil Numbers (UPNs) in the NPD as at 28/12/2015 was 19,807,973. This covers pupil records since 2000.'

1.2 Data releases from the Department

The DfE publishes online a register of third-party recipients to whom it has released data since 2012 through its own application and approvals process (DMAP). Of the registered 462 releases of identifiable data that went through the DMAP in 2012-2014, 53 were aggregated data. All others are individual level. Recent update shows 650+ releases (by end of 2015).

1.3 Data recipients from the Department

In addition to requests for use in public interest research from academic institutions and bodies, data have been released to commercial companies, charities and journalists. Recipients of sensitive identifying personal data include national papers and television. A Freedom-of-Information request shows not all releases are publicly documented. Since 2012 data were given to the Home Office 18 times, and the Police made 31 requests.

2 Methods

We set out to make a preliminary qualitative assessment of awareness in school staff, parents and young people about the NPD, asking them what they know about how children's data collected in school and its use beyond state education. These results could be seen as a pilot for a broader engagement in how the public relate to information and NPD data, and its use by others.

2.1 Responses gathered

In asking school staff about when they last received or made an update to their own privacy policies we encountered consistent difficulty asking about it, as none were familiar with the concept or uses from the NPD. In this atmosphere we promised anonymity to schools and staff in the publication of their responses. Students who gave us recorded interviews gave us only their first name, age, and hometown. We did not ask for contact details to re-contact. We focussed on questions of awareness of data existence and its uses, and asked young people about attitudes to control of their data.

2.1.1 Schools - talking about their pupil data

From a list available online of all state sector schools, and 100 asked, we had replies from 30 primary and 45 secondary schools in Dorset, East and West Sussex, in London, Oxfordshire, and Yorkshire. We selected a spread between rural and city schools, those under Local Authority or academies. We did not ask Free Schools.

At first, we had asked the 75 participants by email to complete an online survey of 5 questions, using Survey Monkey. The survey failed to gather any responses. When we asked schools about it by email or telephone, respondents said it was because they had never heard of the NPD and felt unable to give an informed opinion.

We then instead asked for concrete copies of privacy notice documents via FOI and comment. Privacy policies returned demonstrated a wide variety of wording and consistent gap in communicating NPD use.

2.1.2 Education practitioners questions and answers

We interviewed 100 teaching or affiliated school staff face-to-face at three education events in spring 2016. None were aware of the NPD. Most were aware of the school census but did not know who see pupil data outside school or the Local Authority. Some suggested only statistics are shared outside their school or at national level. Ten staff from Independent schools asked at the Festival of Education, in June, were also unaware.

2.1.3 Parents questions and summary answers

We interviewed 100+ parents face-to-face in November 2015 at the Mumsnet Blogfest in King's Place, NW London. These were parents of children, in education in England, 90% in state education. We discounted 2 home schooling. They came from across England. No one had heard of the NPD or knew that named identifiable data were released beyond school for use by third parties. All were surprised that commercial businesses and journalists could access data. Comments ranged from significant questions of trust, to a lack of concern 'as long as they've not done anything wrong with it.'

2.1.4 Young people questions and their answers

We gathered interviews over two separate hours on two days in May 2016 at the University of Sussex with 25 individuals under 35, only if they had been to school in England, and in different parts of the country. Six agreed to recorded statements. We introduced the idea of the NPD. None had heard of it. We explained that the data has been opened up to third parties since 2012, the approvals process, rules for use, and the wording of the legislation and permitted uses:

"persons who, for the purpose of promoting the education or well-being of children in England are—

(i) conducting research or analysis, (ii) producing statistics, or (iii) providing information, advice or guidance, and who require individual pupil information for that purpose."

We then gave them an A4 guide card and questions can be viewed online. Comments from interviews include:

Catherine, 21, from Gloucestershire: *"Parents and pupils should have access to their own data and should know who else has it. I don't think anyone else should have access to the identifiable data without consent."*

Ben 26, from Reading: *"I don't think commercial businesses should have access to student data. You have not necessarily been exploited, but definitely used."*

Johann 18, from Paris (completed A-levels in England): *"I'm not surprised my data is used by others, probably some of it is used for good causes, but we should know who has it […] we should define our privacy (not the government) and they should ask us before they use it for anything we don't expect."*

Ruby 28, from Newcastle: *"I'm surprised to hear my school data could be used outside schools without my consent. It's a personal thing and can affect lives."*

John 30, from UK: *"I've never heard of the National Pupil Database. I'm really surprised, it's a bit weird. I don't think anyone should have it unless it's to do with my education. We should definitely be asked."*

Steph 19, from London: *"In school I remember being told to do biometric fingerprints for buying lunch. We had no idea what it would be used for and I've no idea if they ever delete them. Parents should be asked for consent.As pupils get older we should decide ourselves."*

Note: It is outside the scope of our own work but we refer to data gathered by two organisations on biometrics in schools. Understanding children's experiences of biometric data (Fingerprinting and RFID) collection and its impact has potential implications for the use of health surveillance data, and willingness to participate in future research. The full national extent of this technology in schools are unknown. (Big Brother Watch, 2014 and Biometrics in Schools, 2010)

2.3 Summary findings

Results show that schools, staff, parents and pupils are surprised to learn identifiable personal data are handed out to third parties at national level. Pupils whose data are in the database and who left school before the 2012 legislative changes of purposes may never have been informed. The updated DfE privacy notice template in May 2016 was the first to contain a direct link to the third-party-recipient register. Our research indicates 2012-13 legislation changes on uses of the data have not effectively reached schools and the DfE fail to effectively communicate any releases of data and the purposes of use to parents and pupils, using the latest DfE suggested or older privacy notice templates.

3 Public Trust

Measures of public acceptance for data use in bona fide academic research in the public interest, and differences in the levels of trust that people attribute to different settings and organisations, were made in The Royal Statistical Society's Data Trust Deficit, with Lessons for Policy Makers (2014). This included views from people aged 16-75 on the use of their personal data in datasets within government. These findings were similar to those from the ESRC Public Dialogues on Using Administrative Data in 2013; and young people, age 14-19 asked in 2010 by The Royal Academy of Engineering (Paterson, L. and Grant, L. eds., Privacy and Prejudice). Few have high trust that government has their best interests at heart using personal data. This improves for anonymous data and non-commercial use.

3.1. Trust levels in young people age 14-19

Young people asked in the 2010 study conducted by The Royal Academy of Engineering (Paterson, L. and Grant, L. (eds) in outreach work supported by three Research Councils and Wellcome, discussed attitudes towards the use of medical records. Questions centred on privacy, and data getting into 'the wrong hands'.

The report concluded: *"These questions and concerns must be addressed by policy makers, regulators, developers and engineers before progressing with the design, development and implementation of EPR record keeping systems and the linking of any databases."(p40)*

Trust in use of their confidential health data was affected by understanding data security, anonymisation, having autonomy and control, knowing who will have access, maintaining records accuracy, how will people be kept informed of changes, who will maintain and regulate the database, and how people will be protected from prejudice and discrimination [through use of their data].

4 Legislation and its implications

Changes in 2012-13 permitted the release of individual data and amended section 114 of the Education Act 2005, section 537A of the Education Act 1996, together with the 2009 Prescribed Persons Act. For detail of the changes before 2007 see *Children's Databases - Safety and Privacy* (Anderson, R., et al. 2006 pp112-115).

4.1 Data Protection law and implications

Current plans to change legislation on the use of public data (Digital Economy Bill 2016 and the EUGDPR) will affect policy and practice and need attention in the UK at the time of writing. To understand the changes, an understanding of current practice is also necessary, specifically of Schedule 2 and 3 and Paragraph 9 of the Data Protection Act and Processing of Sensitive Personal Data Order. (ADRN, legal framework, 2015)

4.1.1 Data use must have been fairly processed

The law can be used to enable safe public interest research for social good, and its limitations are intended to set a high bar for protection of individuals' rights. Use of data in research does not mean Data Protection laws can be disregarded simply because data are deidentified or uses are 'research' and have applicable exemptions.

"Section 33 does not, however, give exemption to the remaining data protection principles.[…] Researchers wishing to use personal data should be aware that most of the data protection principles will still apply (notably the requirement to keep data secure) […] personal data to be fairly and lawfully processed still needs to be adhered to, even if the 'research exemption' applies." (ADRN, legal framework)

This fairness obligation was made explicit again for public bodies in the *Court of Justice of the European Union case (C−201/14)* It ruled the public must be informed when public bodies share their data and why.

4.1.2 Data purposes as foundation of use

The Supreme Court July 28, 2016 ruling on the Scottish Named Persons data sharing plans for children reiterates Data Protection requirements that personal data must be

"*collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes*."

Our evidence indicates NPD purposes are incompatible with those that parents and children give personal data to schools, namely, for the direct purposes of education. The DfE tells schools that they need not ask for consent. The legal foundation that the assumption rests on, the fair processing of collection, and from that the social contract for research, is a data collection privacy notice for parents and pupils. The Department approach is that schools are responsible for fair processing. (Parliamentary question 42842, July 2016).

The school census and early years census collection in 2016-17 intends to expand the quantity of personal data extracted to the NPD to include country-of-birth. (See 1.4 and online for our summary of the changes.) The collection forms we have seen for country-of-birth, and school census, fail to state purposes, or gets them wrong. Scope expansion of uses and users since collection, is now problematic in the National Pupil Database.

This is not aligned with future GDPR obligations. It may fall short regards the UN Convention on the Rights of the Child, Article 16, and Human Rights Act Article 8.

5 Legislative change lies ahead

The Digital Economy Bill will come to Parliament in the autumn of 2016 for debate on significant changes in the handling of all public data. These changes affect both the secondary use of data in anonymised and deidentified use by statistical bodies and accredited researchers, as well as broader access to identifiable data for secondary uses by government at different levels, its agencies and commercial third parties. This offers significant potential for both opportunity and risk to public interest research. The same datasets will be used for multiple purposes, by different users, and share a common foundation in public trust.

5.1 Will legislative changes underpin trust?

Expanding scope use of identifiable data by government is likely to result in further unexpected outcomes for individuals from unseen, processing in the areas of debt such as student loans, fraud and targeted public services (i.e. 'Troubled Families') in which stigmatisation from application, or where the 'freedoms, rights, or interests' of the individual are contrary to those of 'the State'. Conflating new statutory gateways giving access via Trusted Third Party for research and statistics, and also broadening access to all Birth, Marriage and Death civil registration records for wider government use, or to commercial energy companies for example, means increased scrutiny in 2016-17 of all secondary data uses.

5.1.1 Tools to help: Can a Data Science Ethics Framework strengthen public trust?

The new government Data Science Ethics Framework included some public engagement (Sciencewise, 2016). In order to see whether the new framework would help avoid past problems, the questions asked in the ethics framework can be assessed against past programmes. Campaign group medConfidential did this in August 2016 in a blog to demonstrate that the same issues with care.data would reappear if using the new framework. The questions remain how useful this new data ethics framework will be and whom it is designed to serve.

5.1.2 Tools to help: The European Data Protection Supervisor toolkit for policy makers (June 2016)

The EDPS hopes to better equip policy makers and legislators responsible for preparing and scrutinising measures that involve processing of personal data, and which are likely to interfere with the rights to privacy and to data protection and with other rights and freedoms laid down in the Charter of Fundamental Rights of the EU. Case studies highlight legislation which could be of interest in UK practices.

5.2 EU General Data Protection Regulation

The regulation will take effect in all European Union member states in May 2018 and at the time of writing, it appears this will include the UK. Our data infrastructure, data already collected, and all new data collections will need to align with these requirements, particularly the need to consent children adequately, *'freely given, specific, informed and explicit,'* and aligned with Article 6. The notions of data portability, data processor liability, strong provisions on profiling, and enhanced transparency provisions including the public's right to make a Subject Access Request are all be of importance.

5.2.1 Recital 26 - what may it mean for the NPD

Recital 26 will no doubt be the subject of debate. Its

intent is to clarify that pseudonymised data are considered personal data. For research in safe settings their use may be upheld in use of trusted third parties.

For current government department releases of identifiable data it suggests the necessity of an entirely different approach, away from identifiable releases into-the-wild to recipients, and moving towards safe settings.

This may be accompanied by techniques of how to release truly anonymised data as Open Data. Understanding which data in education may be considered Open Data or not, and how access may be made to those which are not, is vital to use data safely.

5.2.2 Recital 40 - impact of incompatible uses

While recital 40 clarifies 'not incompatible' purposes and gives an assurance that the validity of future processing can be based on historical processing conditions, if there is neither consent, nor informed fair processing for the National Pupil Database, on what legal basis will historical data continue processing?

The compatibility of purposes is also necessary to ensure future users are who the data subject expect to be, if their data was collected in the past. Can access to sensitive, identifying or pseudonymous personal data continue to NPD recipients such as consultancies or tutor-pupil matching services, as 'research' purposes?

6 Infrastructure to future-proof UK public interest research data

Part of our public data infrastructure is founded on what types of data we make available to others, whether closed, open or lie on the spectrum in between.

Data infrastructure is as vital to the digital revolution as our transport infrastructure was to the industrial revolution. When data infrastructure flourishes […] we will receive better services, and our environment, our economies and our societies will be improved. (Open Data Institute, 2015) Does the inconsistent infrastructure we have today, work well for our different data needs?

6.1 Trusted infrastructure for NPD needed

If access of the NPD data by today's broad range of third parties were only in safe settings there may still be arguments over who would be considered prescribed persons and who are not qualified 'safe' users, by the standards of the UK Data Service for example. There could also be debate over the extent of data retention rather than deletion, and data minimisation.

However the foundation of good data practices in the data infrastructure in the UK, the physical infrastructure of safe settings, practices following UKAN anonymisation techniques, and accreditation of safe researchers may be the only principles that enable safe and trusted public interest research using population-wide data. Data are accessed through these infrastructures by accredited safe users today. Other users and uses via the DfE undermine this.

Our national pupil data must be made safe if our 'world-leading data resources for social and economic research', should continue to provide 'a huge opportunity to address some of the most pressing challenges facing society,' (ESRC, 2016) opportunity to explore impact, and to hold policy-makers to account.

7 Consensus for change exists

The CMA report (June 2015) on consumer data, highlighted that to secure the benefits of data, people should know when and how their data is being collected and used and decide if and how to participate-.

Policy makers agree. Baroness Kidron said in the House of Lords in January 2014 (Hansard) we should have a regulatory framework that protects young people from the routine collection of their data, that is stored and sold in perpetuity without any recourse. The House of Commons Science and Technology Committee 2014 in their report, Responsible Use of Data, said the Government has a clear responsibility to explain to the public how personal data is being used. Their Big Data Dilemma 2015-16 report, (p9) concluded "*seeking to balance the potential benefits of processing data (some collected many years before and no longer with a clear consent trail) […] is unsatisfactory left unaddressed by Government and without a clear public-policy position.*"

Conclusion

Disparity between government departments and safe research data handling infrastructures, mean inconsistent policy and practices exist in parallel. Secure handling is key to public trust, poor practices jeopardise this and risk data misuse and potential resulting harm.

Consideration must be given beyond the legal requirements to the compatibility of different types of users, and compatibility of users purposes to meet public expectations, if trust in 'the public interest' use is key to securing a social license for research, with no surprises.

Lack of public awareness about data use from pupils in England, and failure to adequately address consent and fair processing are weak foundations for any secondary uses of data collected for direct purposes. There is a consensus that people have the right to know who is holding their data, what their information is used for, why, and whether data are being copied, sold or traded.

To ensure quality data continue to be available for public interest research, a consensual, trusted relationship needs built between data subjects and controllers. If uses across the data spectrum are to best serve our public interest needs, then consistent legal, safe and transparent policy and practice are needed across the data infrastructure, underpinned by accountability, to support public data fit for the future. Fictional future scenarios and 'DNA-chips' may move surprisingly soon into the reality of public policy making. Change is needed today.

Acknowledgements

We would like to acknowledge all our supporters who lend their time and experience to defenddigitalme.

References

Underachievement in Education, (2014) House of Commons Education Committee http://defenddigitalme.com/wp-content/uploads/2016/08/Plomin_-December-2013_142.pdf

Wellcome Trust Briefing (2015) Ensuring the effective use of patient data. https://wellcome.ac.uk/sites/default/files/ensuring-the-effective-use-of-patient-data-briefing-aug15.pdf

Carter, P., Laurie, G., Dixon-Woods, M. (2015) The social licence for research: why care.data ran into trouble, J Med Ethics 2015;41:404-409 doi:10.1136/medethics-2014-102374

The Digital Economy Bill (2016) consultation https://www.gov.uk/government/publications/digital-economy-bill-part-5-digital-government and responses https://www.gov.uk/government/consultations/better-use-of-data-in-government and draft legislation http://services.parliament.uk/bills/2016-17/digitaleconomy.html [accessed August 15, 2016]

Mundie, C. Privacy Pragmatism, Focus on Data Use not Collection, Foreign Affairs, March/April (2014), Volume 93

Hunt, J. Hansard, 25 February 2014, col.147

New data security standards and opt-out models for health and social care (2016) https://consultations.dh.gov.uk/information/ndg-review-of-data-security-consent-and-opt-outs/consultation/subpage.2016-06-22.1760366280/view

To share or not to share? The Information Governance Review (2013)

care.data Programme Board Paper (2013) Reference: ISCG/005/003 http://jenpersson.com/wp-content/uploads/2016/08/ISCG-005-003_Maternitycaredata.pdf

NPD User Guide (2015) http://defenddigitalme.com/wp-content/uploads/2016/08/NPD_user_guide.pdf pp5 and 19-20

NPD numbers, FOI at WhatDoTheyKnow run by mySociety Ltd. https://www.whatdotheyknow.com/request/pupil_data_national_pupil_databa_2

NPD sample analysis http://defenddigitalme.com/wp-content/uploads/2016/04/DDM_shared_examples_April2016.pdf

Third party release register https://www.gov.uk/government/publications/national-pupil-database-requests-received

Sample requests for data made to the Department for Education on WhatDoTheyKnow run by mySociety Ltd. https://www.whatdotheyknow.com/request/pupil_data_national_pupil_databa including The Times (https://www.whatdotheyknow.com/request/293030/response/723407/attach/5/The%20Times.pdf) The Telegraph (https://www.whatdotheyknow.com/request/293030/response/723407/attach/3/Daily%20Telegraph.pdf) Additional third party releases https://www.whatdotheyknow.com/request/pupil_data_sharing_with_the_poli [accessed August 2016]

Online survey for schools https://www.surveymonkey.com/r/KKTCJML?sm=1JEWPcLVPxr%2fitWPf%2f7U8g%3d%3d

Student interviews http://defenddigitalme.com/wp-content/uploads/2016/08/DDM_podcast_interviews.pdf

Big Brother Watch (2014), report https://www.bigbrotherwatch.org.uk/files/reports/Biometrics_final.pdf and report by Pippa King, Biometrics in Schools for the Protection of Freedoms Bill Committee (2010)

Anderson, R., Brown, I., Clayton, R., Dowty, T., Korff, D. and Munro, E., (2006), Children's Databases - Safety and Privacy

Expansion of School Census 2016 http://defenddigitalme.com/wp-content/uploads/2016/07/DDM_COB_expansion_v1.3.pdf

Administrative Data Research Network legal framework (2015) https://adrn.ac.uk/protecting-privacy/legal/dpa/

Judgment of the Court of Justice of the European Union in the Bara case (C−201/14) http://curia.europa.eu/jcms/upload/docs/application/pdf/2015-10/cp150110en.pdf (October 2015)

Supreme Court (2016) UKSC51 https://www.supremecourt.uk/cases/docs/uksc-2015-0216-judgment.pdf

The European Data Protection Supervisor toolkit for policy makers (June 2016)

The Royal Statistical Society's Data-trust-deficit, with lessons for policy makers (2014)

ESRC (2013) Public dialogues on using administrative data, http://www.esrc.ac.uk/public-engagement/public-dialogues/public-dialogues-on-using-administrative-data/

Paterson, L. and Grant, L. The Royal Academy of Engineering (2010), Privacy and Prejudice: Young people's views on Electronic Patient Records. http://jenpersson.com/wp-content/uploads/2016/08/Privacy_and_Prejudice.pdf

Public dialogue on the ethics of data science in government (2016) Cabinet Office, http://defenddigitalme.com/wp-content/uploads/2016/08/data-science-ethics-in-government.pdf

medConfidential (2016) Data in the Rest of Government — the Cabinet Office Data Programme https://medconfidential.org/2016/data-in-the-rest-of-government-the-cabinet-office-data-programme/ [accessed August 2016]

The Open Data Institute (2015) How will the future affect data infrastructure? http://theodi.org/blog/how-will-future-affect-data-infrastructure [accessed August 2016]

UK Data Service accredited researchers https://www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab [accessed August 2016]

CMA report (2015) Commercial use of consumer data https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/435817/The_commercial_use_of_consumer_data.pdf [August 2016]

The House of Commons Science and Technology Committee 2014 Report, Responsible Use of Data http://www.publications.parliament.uk/pa/cm201415/cmselect/cmsctech/245/245.pdf

The Science and Technology Committee Big Data Dilemma Report (2015-16) http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf