The Who, What and Why: An Analysis of Personal Data Transparency Notices in the UK

Reuben Binns[*], David Millard[*], Lisa Harris°

[*] *Web Science Doctoral Training Centre University of Southampton Southampton, UK*
° *Web Science Institute University of Southampton Southampton, UK*

Abstract.
Data protection laws require organisations to be transparent about how they use personal data. This article explores the potential of machine-readable privacy notices to address this transparency challenge. We analyse a large source of open data comprised of semi-structured privacy notifications from hundreds of thousands of organisations in the UK, to investigate the reasons for data collection, the types of personal data collected and from whom, and the types of recipients who have access to the data. We analyse three specific sectors in detail; health, finance, and data brokerage. Finally, we draw recommendations for possible future applications of open data to privacy policies and transparency notices.

Keywords: privacy, data protection, personal data, transparency, web, open data

1. Introduction

The use of personal data has become one of the most important issues of the digital age. Regulators, policy-makers and consumer advocates have long argued for transparency from the public and private entities who gather and use this data. Transparency is a core principle at the heart of several foundational privacy and data protection frameworks, and continues to inform new regulations and international instruments[1]. Although transparency alone may be insufficient, it is seen as a necessary precondition to achieving privacy goals (Gutwirth & DeHert, 2006). In theory, it helps regulators, advocates, researchers and others to monitor and analyse privacy-related practices, guiding their strategy and further action. Ultimately, transparency also aims to empower privacy-conscious individuals, whether directly or through an intermediary, to make more informed choices about whom to trust with their data (Egelman, Tsai, Acquisti, & Cranor, n.d.). Transparency is therefore a prerequisite for a functioning 'market for privacy', where companies compete on their privacy credentials in order to attract privacy-sensitive consumers.

But despite broad support for the principle and purpose of transparency, there has been less agreement on how best to achieve it. Effectively and efficiently recording and publishing what organisations do with personal data, and why, has proven difficult. A significant body of research has built up around the design and testing of improved transparency mechanisms. At the same time, many studies attempt to use existing mechanisms – which principally come in the form of 'privacy notices' – to analyse the policies and practices of organisations regarding personal data. But these studies face significant barriers which limit their potential depth and scope.

This paper addresses both the design of privacy notice transparency systems, and the analysis of their content. We present an analysis of a previously unstudied source of standardised privacy notifications, the UK Register of Data Controllers[2], which contains notifications made to the UK Information Commissioners Office (ICO) by around 350,000 organisations over an 18 month period. Our aims are to generate a broad overview of the landscape of personal data use by UK organisations, and bring new evidence to bear on some topics of pressing public concern and previous study, namely the

[1] See, for instance, the 'Openness Principle' in (OECD, 1980)

[2] Available to search at http://ico.org.uk/esdwebpages/search

collection and use of personal data by health providers, financial services, and data brokers. The results of the analysis are followed by considerations and recommendations for the creation of such transparency systems.

2. Background

In order to provide context, the remainder of this introduction briefly outlines the history and current status of privacy notice transparency systems, some developments which have been proposed, as well as related examples of mass analysis of privacy notices.

Transparency has so far in practice been implemented through the use of notices, often published by organisations as 'privacy policies' to be included alongside their terms-of-service and end-user license agreements. The common practice of producing lengthy and legalistic documents means that few consumers read or understand these policies. A study of online privacy policies estimates that the average U.S. Internet user would have to spend 244 hours per year reading the privacy policies of all the websites they visit during that year, suggesting that the cost of being informed may well be too high for any individual (McDonald & Cranor, 2008). The length of these documents is also a barrier to academic research and regulator investigations into organisations' stated privacy practices. Several commercial and non-profit organisations have attempted similar work, classifying and rating privacy policies on behalf of consumers. But manually parsing the mass of policies is a time-consuming task, which has limited the coverage and effectiveness of these efforts.

While privacy notices have received most of the attention in the transparency debate, certain jurisdictions also maintain an alternative scheme of public registers[3] This approach involves mandatory disclosures by organisations to a regulatory authority, detailing what data they collect, who they share it with,

[3] Most E.U. member states have such registers, but exceptions include Germany and Sweden.

and why. This information is then gathered in a national register of organisations' personal data practices, which is made available to the public. This system – implemented in most EU member states – is generally held in low regard, with the EU Commission describing it as an 'unnecessary administrative requirement'[4]. At the time of writing, only eight of those member states with national registers appear to have public websites from which they can be searched, which are of varying quality and usability. Anecdotal evidence suggests that public awareness of these public registers is limited to a small number of data protection specialists, and those who do attempt to use them for transparency purposes find they have low usability and are inconvenient[5]. Given their perceived lack of utility, it is unsurprising that the new draft proposal for a General Data Protection Regulation (henceforth 'GDPR')[6] dropped mention of such registers altogether.

Both privacy notices and national registers fall short of the kind of transparency system that would be required for meaningful oversight, monitoring and analysis by regulators, researchers and consumers. But despite the problems with the existing measures, policy-makers continue to emphasise the need for transparency more than ever. At a 2013 international meeting, privacy and data protection commissioners released a statement on transparency, recognising that:

> “Effective communication of an organisation’s policies and practices with respect to personal data is essential to allow individuals to make

[4] 'Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses' – European Commission, Press Release http://europa.eu/rapid/press-release_IP-12-46_en.htm

[5] See documented complaints made by an individual attempting to use the UK's online register: https://www.whatdotheyknow.com/request/non_notification_team

[6] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)

> informed decisions about how their personal data will be used and to take steps to protect their privacy and enforce their rights."[7]

With the desire for transparency greater than ever, there is renewed enthusiasm for new approaches to effectively record and publish organisations' policies.

Some propose standardised, short, simplified and graphical notices. The U.S. Department of Commerce has proposed guidelines for short form notices to describe third party data sharing (NTIA, 2013), while the proposed GDPR advocates a 'nutrition label' style approach, complementing traditional notices with standardised and required fields represented by a set of common simple visual icons that would become familiar to consumers over time[8]. This approach has also been explored by a number of non-profits and consumer-oriented companies[9] and more recently was the subject of an initiative by the ICO to develop 'privacy seals'[10]. Similar initiatives aim to encode privacy policies into machine-readable XML formats[11]. A common theme in all of these proposals is the idea that by standardising and digitising organisations' disclosures of how they collect and use personal data, this information can be aggregated, accessed, compared and analysed en mass by regulators, consumer advocacy groups, intermediaries or individuals themselves.

As has been noted elsewhere (Lorrie Faith Cranor, 2012), this is not an entirely

[7] At the 35th International Conference of Data Protection and Privacy Commissioners (2013), Resolution on openness of Personal Data Practices, 3–5.

[8] See article 13(a) of the Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation).

[9] See for instance, Mozilla Icons project (https://wiki.mozilla.org/Drumbeat/Challenges/Privacy_Icons), ToS-DR, PrivacyScore

[10] "ICO to launch privacy seals scheme 'within the year'", DataGuidance 27/03/2014 http://www.dataguidance.com/dataguidance_privacy_this_week.asp?id=2258

[11] See proposals from TRUSTe, a web certification body (http://www.truste.com/blog/2010/09/14/more-on-the-problem-with-p3p/) , and the Internet Advertising Bureau's CLEAR Ad Notice project (http://www.iab.net/clear).

new idea. The current policy proposals have strong parallels with ambitious efforts in previous decades to create large-scale systems for the transparent use of personal data. Amidst enthusiasm for new measures, there is a danger of reinventing old (failed) solutions, unless we learn from past attempts. These previous initiatives, their promises and failures, could be instructive in setting the context and guiding the development of new transparency systems.

Perhaps the most significant long-term effort in this vein came from a series of initiatives which began with the Platform for Privacy Preferences (P3P)[12] in the mid-1990's. Early proponents described a system whereby privacy policies could be encoded as structured data. This data could be 'understood' by web browsers and other software agents, which could then automatically negotiate with websites on behalf of users according to their privacy preferences – ultimately creating a 'market for privacy'. By the time the World Wide Web Consortium (W3C) specification for P3P was approved in 2002, the negotiation features had been dropped, however a standard for rendering privacy policies in machine-readable XML format remained. The standard had a number of early adopters including news websites, search engines, ad networks, retailers, telecommunications companies and government agencies (Lorrie Faith Cranor, 2013). The hope within the web standards community, in particular amongst proponents of the 'semantic web'[13], was that a significant proportion of organisations would independently adopt the standard, thus creating a decentralised database of organisation's privacy practices. If successful, such a system could be intelligently queried and analysed en mass, thus helping the activities of regulators, intermediaries and consumers.

Perhaps inspired by this vision, the standard was spurred on by regulators in

[12] http://www.w3.org/P3P/

[13] The semantic web vision is to turn the human-readable content of the existing world wide web into a machine-readable 'consistent, logical web of data' (Berners-lee, 2004). For an example application of P3P in the semantic web, see (Gandon, 2003)□.□

the U.S., principally the FTC. The standard was initially envisioned as a framework for consumer-focused tools, but the FTC also noted the potential of P3P for use in their own investigations and enforcement actions. In 2001, the FTC incorporated P3P data into their annually commissioned surveys of website privacy policies (Milne & Culnan, 2002), which were conducted in order to investigate organisations' adherence to the FTC's 'Fair Information Practices'. One of the policy recommendations arising out of these studies was to encourage businesses to adopt the emerging standard for their websites. This would make future longitudinal analysis of privacy practices more effective and comprehensive due to the potential for automated analysis. The gradual adoption of this technology by websites in the following years did result in such work. The first detailed and large-scale analysis of the policies of P3P-enabled websites was subsequently conducted in 2003. It investigated the types of data collected, the uses to which it was put, and the types of recipients the data is shared with (Byers, Cranor, Kormann, Ave, & Park, n.d.).

Unfortunately, further studies like this were hampered by the decline of the standard. When a modified version of the (Byers et al., n.d.) study was repeated in 2006, it was found that the proportion of P3P-enabled policies containing errors had increased. Despite evidence of their increased usability (Kelley, Cesca, Bresee, & Cranor, 2009), and backing from regulators, the use of standardised P3P privacy notices began to decline. A 2007 study indicated that the level of P3P adoption in 2005 was low (8.4%), and showed that adoption had remained stagnant since 2003 (Beatty, Reay, Dick, & Miller, 2007). By 2010, P3P 'compact policies' (shortened versions of full P3P-enabled privacy policies) were even found being used to mislead rather than inform users (Leon, Cranor, Mcdonald, & Mcguire, 2010). As of April 2014, support for the standard has been dropped by all the major web browsers apart from Microsoft Internet Explorer[14]. In a retrospective policy piece, a former developer of the

[14] IE blocks third-party browser cookies by default if they do not have P3P policies (see

P3P standard suggests that it failed because it was too complex for websites to translate their privacy policy into the P3P format (Schwartz, 2009).

Other policy languages have been designed to supersede P3P, but none have achieved significant adoption as yet[15]. They may face a 'network effects' problem, in that the positive effects of standardisation only emerge once a significant portion of organisations/websites have adopted the standard (Tsai, Egelman, Cranor, & Acquisti, 2010). Therefore, the initiatives struggle to get off the ground as their full benefits are hard to demonstrate. Similar studies of privacy policies by academics and regulators have continued in the absence of P3P or other standards[16], but their scale and reach is limited by the fact that policies must be parsed manually before any analysis can be done.

Meanwhile, the alternative transparency system of public registers has received far less attention than P3P and its various relatives. Perhaps the earliest reference to the public register model in international privacy and data protection frameworks can be found in the 1980 OECD privacy guidelines[17], in the detailed comments elaborating on the 'Openness Principle'. The guidelines note that openness is a pre-requisite for individuals to exercise their right to access and challenge personal data. One of the suggested means to achieve such openness is through the 'publication in official registers of descriptions of activities concerned with the processing of personal data'[18]. The OECD guidelines formed the basis for many subsequent national privacy and data

http://www.techrepublic.com/blog/software-engineer/craft-a-p3p-policy-to-make-ie-behave/)

[15] See, for instance the Primelife Policy Language (Vimercati, Paraboschi, & Pedrini, 2009). The 'Do Not Track' standard – in which a preference/policy regarding online 'tracking' can be communicated between a client and a server – can also be seen as a (minimally expressive) descendent of the P3P standard (see http://www.donottrack.us).

[16] See, for example, the Global Privacy Enforcement Network's 'privacy sweep' investigation of website privacy policies available at https://www.priv.gc.ca/media/nr-c/2013/bg_130813_e.asp

[17] See the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (OECD, 1980)

[18] Ibid, 'Paragraph 12: Openness Principle' (OECD, 1980)

protection regulations and frameworks, with the result that requirements for national public registers are in place in many countries, particularly in the European Union (which in turn has had a significant influence on the development of data protection laws elsewhere (Greenleaf, 2012).[19]

The idea of a centralised public register of organisations' privacy notifications has both similarities and differences to that of a decentralised, machine-readable corpus of privacy notices. Both systems evolved separately, developed by different communities, yet there are similarities in their original visions. Both aim at a comprehensive resource of standardised privacy notifications. Their initial implementations certainly differed, with P3P conceived as a decentralised, data-driven system from the outset, and the public register as a highly centralised, analogue resource, conceived of before the advent of widespread personal computing. But in more recent years, many national public registers have been published online[20], and in some cases made available as machine-readable open data – in a similar format to the P3P standard[21].

In this form, public registers arguably come closer to the original, semantic web vision for P3P than the P3P initiative itself ever did. They contain highly standardised, complete, machine-readable privacy notices from a wide range of organisations. Indeed, the number of organisations contained within the UK register (~350,000) is comparatively far larger than the number of

[19] There are similarities between public register schemes in operation in Europe and US proposals for registers of data brokers and their activities. The FTC have encouraged "creating a centralized website where resellers would identify themselves and describe how they collect and use consumer data, and the access rights and other choices that consumers have" (United States Government Accountability Office, 2013). A similar public register system covering the use of personal data by government agencies was part of the U.S. Privacy Act of 1974, but was also later criticised for being under-used by ordinary citizens (U.S. White House OMB, 1983, p118)

[20] As well as the UK register which is the subject of this paper, other jurisdictions with online registers include Austria, Belgium, Czech Republic, Ireland, Poland, Portugal, Spain and Serbia.

[21] Both the UK and Poland have stored their register data as machine-readable XML, with fields that correspond almost exactly to some of the standard P3P fields.

organisations with P3P-enabled privacy policies identified in any previous study[22]. In addition, because inclusion in a public register is generally mandatory and enforced by law, the contents of a register are less likely to be biased towards those organisations who would voluntarily adopt a given standard (something which was likely to be the case for P3P).

Noting their similarities - whilst being mindful of their differences - it is possible to draw parallels between the public register data and the corpus of privacy policy data that is the subject of prior studies. Analysis of the register data can therefore be seen as a continuation of the extensive body of existing research into organisation's privacy policies, with the advantages of automation and magnitude (which is lacking in previous studies predicated on manually parsed policies), and completeness (which the P3P studies lack). At the same time, any design insights derived from this analysis are likely to be highly applicable to the various proposals for transparency systems mentioned above.

As well as the aforementioned FTC-commissioned research, numerous other studies have aggregated and manually parsed privacy notices to derive quantitative insights into organisational policies and practices regarding personal data. Such studies usually aim to identify trends in organisations' stated practices, and/or evaluate the notification/disclosure process itself. Since a prime motivation for studying (and regulating) the use of personal data is to further the interests of data subjects and society at large, this research is often driven, at least partly, by public attitudes and concerns.

Our general analysis extends existing research by providing a broad overview of the reported uses of personal data across the whole range of sectors and uses. This is complemented by in-depth analyses of three specific uses – trading of personal data, financial services and health provision – each of which have

[22] The largest number of P3P-enabled websites found in any of the prior studies identified in our literature search was 14,720 (Lorrie F Cranor, Egelman, & Sheng, 2008)

been the subject of sustained interest among researchers and in the public eye:

Trading of personal data: Organisations have come under increasing scrutiny over the buying and selling of personal data in recent years. The 'data broker' industry – where personal data is collected and re-sold – has been the subject of investigations by regulators and media organisations[23]. This is also a theme arising in multiple studies of consumer concerns, where it is frequently expressed in terms of 'third parties' with whom data may be shared. In a qualitative study of UK citizens, it was found that an 'unspecified reference to 'third parties' unsettled participants and helped feed concerns that after the transaction there would be a number of uses of their information over which they could have no control'(Bradwell, 2010). An E.U.-wide study found that of the 54% of citizens who were aware of organisations selling their personal information to third parties, only 35% found the practice ethically acceptable (Brockdorff & Appleby-arnold, 2011). In the following analysis, we examine the extent and nature of this practice as compared to other practices, using the pre-defined register category of 'Trading / Sharing in Personal Information'.

Financial Services: Previous research has examined the extent to which organisations in a particular industry or context actually differ in their practices, in order to assess whether there is the possibility of meaningful consumer choice and a differentiated market for privacy (Lorrie Faith Cranor, Idouchi, Leon, Sleeper, & Ur, 2013), (Bonneau & Preibusch, 2010). (Lorrie Faith Cranor et al., 2013) took advantage of a widely implemented standard for privacy notices adopted by 3,422 US financial institutions. In this rare instance of a relatively successfully adopted standard notification format, large-scale empirical analysis of their privacy practices was possible. The authors found significant variety in bank's practices, as well as some evidence of self-contradiction and non-compliance by some institutions. Using data on UK

[23] For instance, the FTC - see (Federal Trade Commission, 2013) and the Wall Street Journal's 'What They Know' series (online.wsj.com/public/page/what-they-know-digital-privacy.html)

banks and other organisations providing financial services, we similarly investigate whether there is homogeneity or the possibility of meaningful consumer choice for UK consumers.

Health services: In a qualitative survey of attitudes towards privacy and health information, national health service patients in the UK regarded health data as a special category worthy of particular concern(Wellcome Trust, 2013), a finding that is supported in earlier E.U.-wide quantitative studies (Brockdorff & Appleby-Arnold, 2011). In February 2014, UK government proposals to share medical data gathered from general medical practitioners under the care.data scheme raised controversy and debate about the risks of sharing health data[24]. We present a profile of data use by organisations engaged in health administration and services, in order to provide context to concerns about these kinds of practices.

These specific analyses are presented alongside the analysis of data collection in general (i.e. for all purposes) for comparison. This shows, for instance, whether different kinds of data are more often collected, or whether certain kinds of data subjects are more often involved, in the context of health, finance, or trading, than in the general case. Previous studies have been unable to provide such a comparison, because they are generally limited to particular sectors (e.g. financial companies or social networking sites), with sample sizes that are both small and unrepresentative. By presenting a comprehensive, representative, cross-industry overview of organisations privacy practices, we aim to situate a particular sector or practice in its broader context.

3. Methodology and Data Source

The source of the data in this analysis is the United Kingdom Information

[24] See 'Care.data: How did it go so wrong?', BBC News, 19 February 2014, (http://www.bbc.co.uk/news/health-26259101)

Commissioner's Office (ICO) public register of data controllers. Data controllers are defined as 'a person who (either alone or jointly or in common with other persons) determines the purposes for which and the manner in which any personal data are, or are to be, processed' (UK Data Protection Act 1998 [DPA], s.1)[25]. 'Data processor' is defined as 'any person (other than an employee of the data controller) who processes the data on behalf of the data controller' (DPA 1998, s.1). Personal data is defined as 'data which relate to a living individual who can be identified (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller', while 'data subject' is defined as 'an individual who is the subject of personal data' (DPA 1998, s.1).

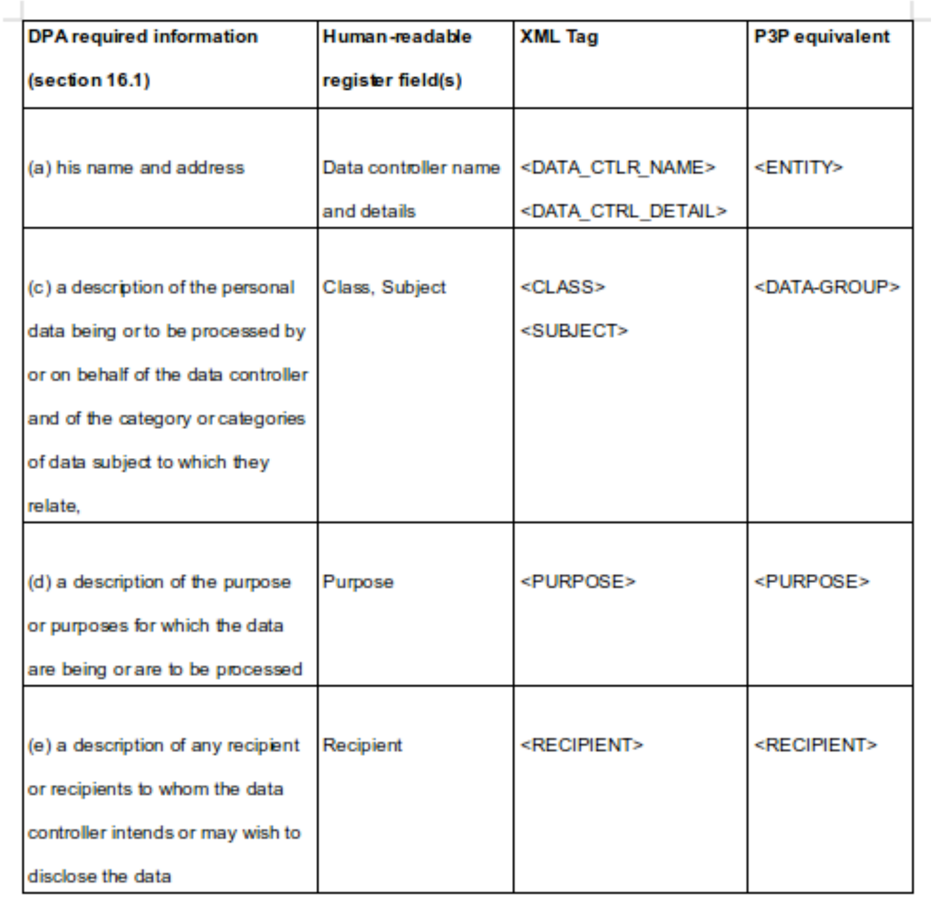
The DPA states that data controllers must contact their national supervisory authority, notifying them of their name and address, purposes of processing, the categories of data types and subjects to whom they relate, recipients to whom the data may be disclosed, and proposed transfers of the data to third countries (DPA 1998, s.16). Furthermore, these notifications should be compiled into a register of data controllers, made available for inspection by any person (DPA 1998, s.19(6)(a)). In the UK, this register is made available to the public to search on the ICO's website, and a regularly updated version of the whole register is available upon request under an Open Government License in a re-usable, machine-readable format. The latter, gathered over an 18 month period, forms the basis of the following analysis.

4. Data structure, extraction and selection

The register is made available in a semi-structured standard data format (XML), and contains fields corresponding to a)-e) of the notification requirements in the DPA (section 16.1). The ICO provide a set of standard

[25] Note that 'person' in this context means 'legal person', and as such could be an organisation or a natural person.

defined purpose types, subjects, classes, and recipients. Data controllers may also describe their activity in their own terms if it is not captured by these standard definitions. In all, three copies of the register were used, from September 2011, September 2012, and March 2013 (unfortunately, data from September 2013 is unusable for the purposes of this study due to changes made by the ICO in April that year).



DPA required information (section 16.1)	Human-readable register field(s)	XML Tag	P3P equivalent
(a) his name and address	Data controller name and details	<DATA_CTLR_NAME> <DATA_CTRL_DETAIL>	<ENTITY>
(c) a description of the personal data being or to be processed by or on behalf of the data controller and of the category or categories of data subject to which they relate,	Class, Subject	<CLASS> <SUBJECT>	<DATA-GROUP>
(d) a description of the purpose or purposes for which the data are being or are to be processed	Purpose	<PURPOSE>	<PURPOSE>
(e) a description of any recipient or recipients to whom the data controller intends or may wish to disclose the data	Recipient	<RECIPIENT>	<RECIPIENT>

Tab;e 1. Comparison of DPA requirements and related data fields

Because the original XML file was too large to query directly, we first parsed the data using SAX, an event-based sequential access parser API for XML[26]. It was then restructured as an SQL database, composed of separate tables for each of the human-readable register fields, along with unique identifiers for each data controller and each 'purpose' instance. This database was then queried to extract relevant portions for further analysis.

5. Analysis

The first stage of analysis was to measure the occurrence of different classes in the dataset. Given that the data is exclusively concerned with categories (i.e. nominal data), in order to subject it to quantitative analysis we measured the occurrence of certain classes. Quantifying the remaining categories (purpose, subject, class, and recipient) reveals the extent to which certain arrangements and relationships exist within organisations. We can then derive conclusions about the extent and nature of data sharing between individuals, data controllers, and third parties, and the prevalence of certain types of personal data, data subjects, and recipients.

Categories with similar definitions (e.g. 'Marketing' and 'Marketing, Advertising and Public relations') were aggregated. Any category with less than 50 instances that could not be meaningfully aggregated into a more prevalent category was discarded. By conducting the same operations on each of the datasets (from 2011, 2012, and 2013), we measure differences in practices over time. The three specific analyses followed a similar procedure with some differences. For the analysis of personal data use in financial services, two subsets of the data were isolated and analysed. One large subset consisted of instances where data was collected and used for the purposes of providing financial services and advice – this included a wide variety of different organisations, not just banks (37,436 distinct organisations in total). A second,

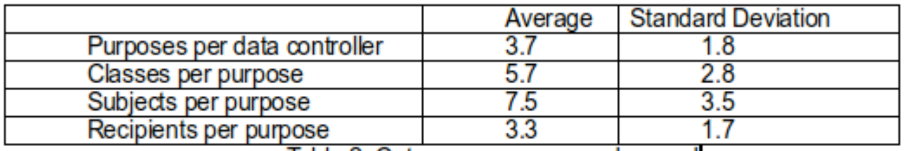
[26] See www.saxproject.org

smaller subset consisted of 98 data controllers whom we independently (manually) classified as 'retail banks'. These samples were then analysed to establish which classes of data were used (e.g. 'Personal Details' or 'Employment'), and which categories of recipients had access to this data (e.g. 'Credit Reference Agencies' or 'Regulators').

6. Results

We found steady growth in overall data collection, the types of data involved, and the types of entities who have access to the data. Each of these fields exhibit a power law distribution with a few very common categories accounting for the majority of the total. The following figures present the general and specific cases side-by-side for ease of comparison.

The total number of data controllers averaged 358,558 across the time period studied, growing by 6.5% from 346,589 in September 2011 to 369,323 in March 2013. The number of purposes (which could also be understood as the total number of distinct reasons for which data is used) exhibited a similar level of growth of 6.3%, from 1,291,075 to 1,371,884. The average number of purposes per controller stayed consistent across the period at an average of 3.72, indicating that while the number of organisations classified as 'data controllers' is increasing, the average number of different types of uses of data per controller remains the same. The standard deviation in number of uses is 1.8, indicating that most data controllers are close to this average. The average number of distinct types of subject, class, and recipients per purpose provide a benchmark for analysis of specific sectors and practices, where averages and spread may differ.



	Average	Standard Deviation
Purposes per data controller	3.7	1.8
Classes per purpose	5.7	2.8
Subjects per purpose	7.5	3.5
Recipients per purpose	3.3	1.7

Table 2. Category averages and spread

The remaining general analysis is broken down by the five fields of 'Purpose' (i.e. why data is collected / used), 'Subject' (who the data is about), 'Class' (what categories of data are collected / used), and 'Recipient' (who is given access to the data). In addition to the total number of entries per category within a field, the prevalence of each purpose category can be calculated in relation to the total number of 'data controller' instances in the entire register, indicating the proportion of organisations engaging in that practice. Similarly classes, subjects, and recipients are expressed as a percentage of the total number of uses (or 'purposes') in the register. This provides a more natural measure, expressing how often a given category appears as a proportion of the total number of data controllers or uses (figures 1-5 express this as percentages).

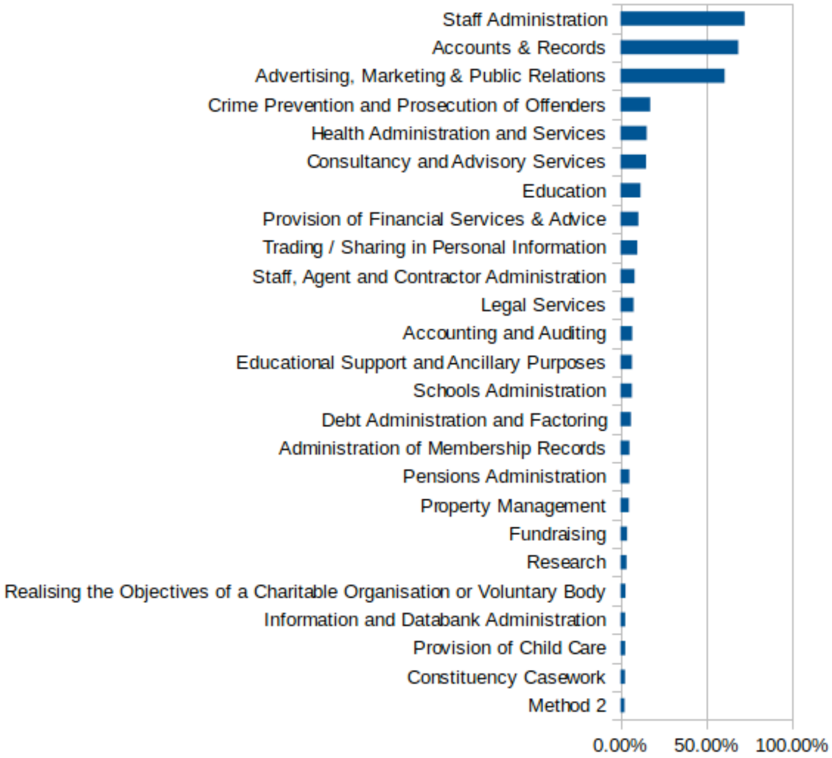
Across all fields, the registrar-defined standard descriptions (i.e. those which are explicitly defined in the ICO's notification handbook given to registrees) featured more heavily than registree-defined descriptions (i.e. those invented by data controllers themselves). The distribution of entries for each of the categories within a field tended to follow a power law distribution, with a few very prominent categories having a high number of entries, and a 'long tail' of more obscure, generally registree-defined categories.

7. Why is data being processed?

The three most common categories in the 'Purpose' field (namely 'Staff Administration', 'Accounts & Records', and 'Advertising, Marketing & Public

Relations') accounted for 54% of the total on average across the period, while the bottom 14 categories accounted for just 13%. Changes between the number of entries for a given purpose category were measured in the 18 month period, with a mean growth of 6% across all categories. While the top 5 categories ('Staff Administration', 'Accounts & Records', 'Advertising, Marketing & Public Relations', 'Crime Prevention and Prosecution of Offenders', and 'Health Administration and Services') grew between 5-10%, the most significant growth was found in more obscure, registrar-defined categories such as 'Provision of Childcare' and 'Provision of Investment Management and Advice'. However, this apparently large change is likely to be amplified due to the relative size of the obscure categories as a proportion of the total.

Figure 1. Purposes for processing personal data (% of controllers claiming)

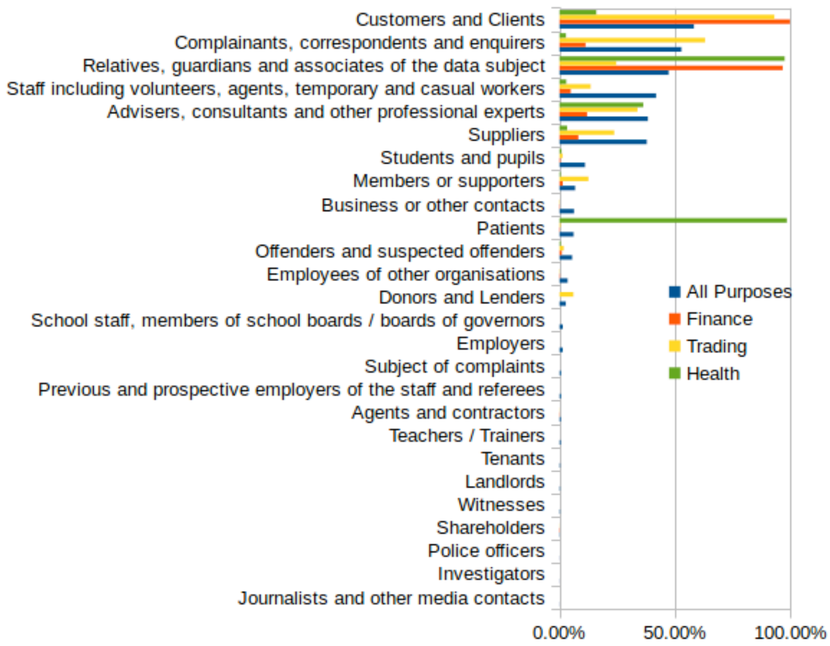


The three purposes which have been selected for further analysis account for a significant minority of all uses. The use of personal data for 'Health administration and services' was listed by 15% of all data controllers on average across the time period (the fifth most common purpose). 'Provision of financial services and advice' and 'Trading / Sharing in Personal Information' were both listed by around 10% of controllers, and were (respectively) the eighth and ninth most commonly listed uses of data.

8. Who is the data about?

The five most common types of data subjects accounted for 85% of the entire field, while the 14 least common accounted for just 4%. The growth for all types of subject was similar to the overall growth in purposes (6.5%). The two categories with the biggest growth were 'Subjects of complaints' and 'Landlords'.

Figure 2. Data subject types (% of all purposes)



In comparing health data to the general case, we find, unsurprisingly, that personal data is far more likely to be about patients and relatives, and far less likely to be about customers and complainants. The trading of personal data appears to mirror this in reverse, involving relatively few patients and relatively many customers.

9. What kind of personal data is used?

The registrar-defined data classes constituted the majority of the categories in the 'class' field, with only five registree-defined classes achieving more than 50 entries. In cases where data is used for healthcare purposes, this often includes sensitive data[27], for instance about an individuals sexual life, which is collected in over 80% of cases compared to just 7.7% across all purposes.

[27] According to the Data Protection Act, personal data in these categories requires special treatment by data controllers. Generally, it may only be processed under special conditions, such as when processing is in the public interest, or when the data subject has given their explicit consent. See the ICO's guidance, available at [http://www.ico.org.uk/for_organisations/data_protection/the_guide/conditions_for_processing]

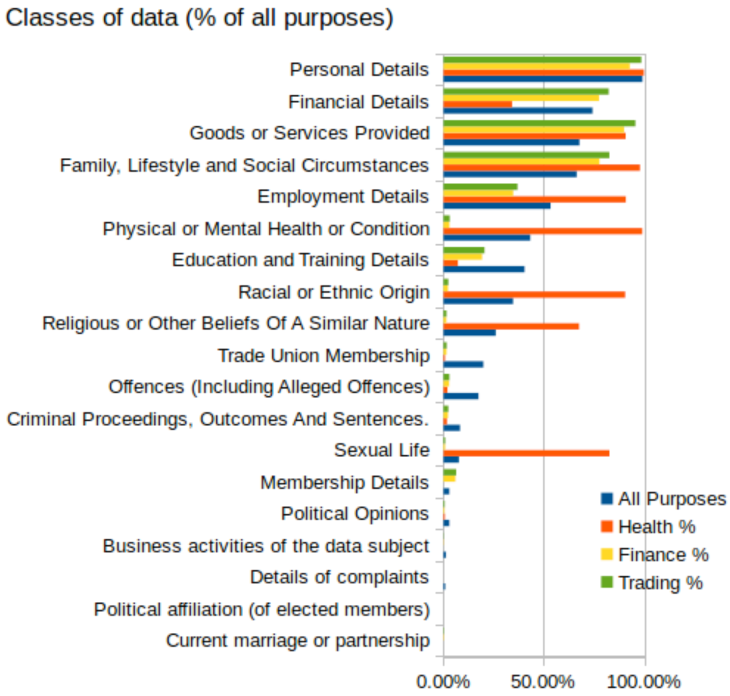


Figure 3. Classes of data

The classes of data collected for the purposes of trading include personal details (99%), goods provided (96%), family and lifestyle (83%), and financial details (82%). These are the kinds of personal information one might expect to be traded, given that they may pertain to commercially useful knowledge like the kinds of goods people might buy or their creditworthiness. However, a small proportion of instances where personal data was traded / shared involved more sensitive kinds of personal data. The ICO lists 8 classes of 'sensitive' data, all of which were 'traded' in the following percentage of cases: Physical or Mental Health or Condition (10%), Racial or Ethnic Origin (8%), Religious or

Other Beliefs Of A Similar Nature (6%), Trade Union Membership (4.6%), Offences (Including Alleged Offences) (4%), Criminal Proceedings, Outcomes And Sentences (1.9%), Sexual Life (1.8%), and Political Opinions (0.7%).

Amongst both retail banks and providers of financial services more generally, the use of personal and financial details, and information about goods and services provided is almost ubiquitous. Unsurprisingly, at least 97% of entries stated using this information. However, there was more variation in practice concerning other types of data. For example, a quarter of retail banks did not list 'Employment Details', and only half listed 'Education and Training Details'.

Growth across all categories was uniform at around +5%, with the exception of 'Details of complaints' reaching a high of +18% growth.

10. Who has access to the data?

By far the most common recipient (or potential recipient) of personal data is the data subject themselves; in fact, out of 1,332,723 uses, the vast majority (92%) give access to the data subject themselves. This is probably due to the fact that under UK data protection law, in most cases, data subjects have the right to request a copy of data held about them (exceptions apply in some cases such as criminal investigations). Average growth across all categories was 5.8%.

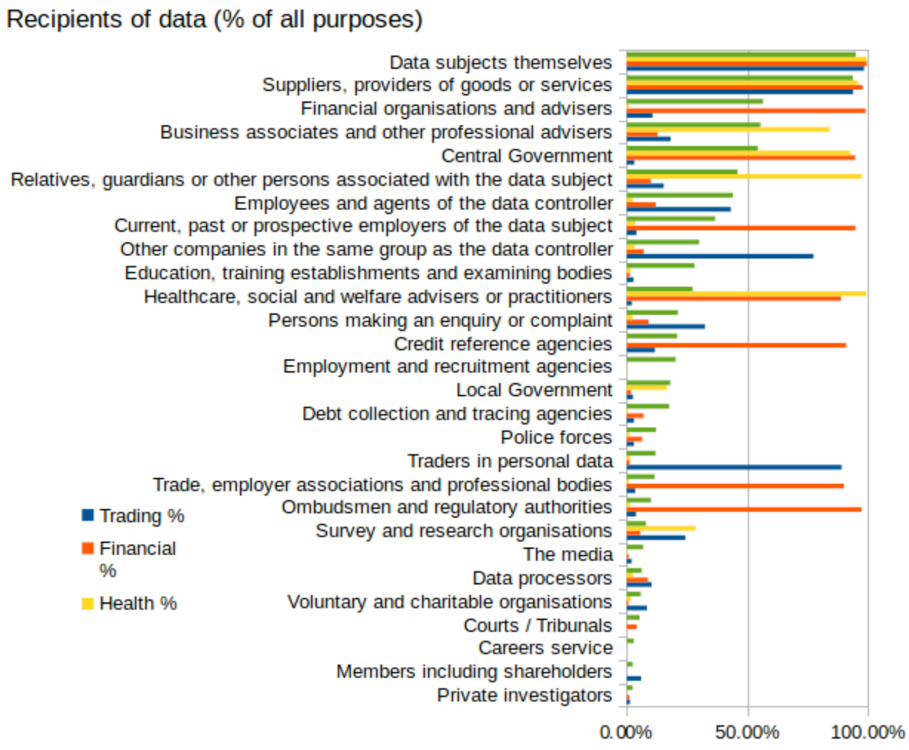


Figure 4. Recipients of data

Overall, in situations where data controllers were 'trading / sharing in personal data', the average number of subjects is 2.7 – much less than the general average of 7.5. However, the average number of recipients was 5.7 – higher than the general average of 3.3. This indicates that when personal data is traded or shared, it is likely to involve only relatively targeted types of data subject, but the data will then likely be shared with a broader than average range of recipients. As with other categories, the most commonly stated potential recipient for this is 'Data Subjects themselves' (100%), as is generally required by law, followed by 'Suppliers, providers of goods and services' (96%) and unsurprisingly, 'Traders in Personal Data' (90%).

Interestingly, the aforementioned classes of 'sensitive' personal data were also being traded/shared with a wide range of recipients. We further investigated the use of sensitive data classes in trading/sharing, finding 134 organisations who state that they trade/share data about individuals political opinions with 'credit reference agencies'. Data about individual's sexual lives is reportedly traded/shared with 'Traders in personal data' by 226 organisations. 'Trade, employer associations and professional bodies' reportedly receive data about individuals' trade union membership from 288 organisations, and their racial or ethnic origin from 182 organisations.

In the case of access to data collected for provision of financial services, the kinds of entities who have access to this data (i.e. those listed as 'recipients') exhibited a similar pattern. Both providers of financial services (displayed in figure 7) and retail banks more specifically almost always gave access to a certain familiar list of entities, such as 'Data Subjects themselves' (as is normally required by law), 'Employees of the data controller', and 'Suppliers and providers of services'. However, there were also some differences between the practices of retail banks and general financial service providers. For instance, while 72% of the former shared this data with 'Data Processors', only 9% of the latter did so. Similarly, giving 'traders in personal data' access to the data was more prevalent amongst retail banks than financial service providers (22% versus 1% respectively). Perhaps surprisingly, this trend appears to reverse when it comes to sharing data with credit referencing agencies, where only 52% of banks share, compared to 90% of financial services providers generally.

11. Discussion

The general analysis here supports the perhaps unsurprising hypothesis that the use of personal data is increasing, in so far as the total number of entries in all fields is growing. However, this conclusion should be accompanied with the following considerations. First, since entries in the register describe the existence of certain data collection, usage and sharing arrangements, the existence of more entries should not be confused with other measures of data use, such as volume (in terms of database entries, queries, or bits). Second, an interesting finding is the growth rate seems to be driven by new data controllers, rather than an increase in the overall counts of purposes, data classes, subjects, or recipients per purpose. In other words, the number and range of uses of personal data by individual organisations does not appear to be increasing, but the total number of organisations registered as data controllers is.

The power law distribution we observed in each field, where a few highly common categories account for the majority of the entries in a given field, is in keeping with previous research. A similar distribution was observed in classes of personal data collected by US banks (Lorrie Faith Cranor et al., 2013) and websites (Culnan, 2000), where a relatively small number of classes account for the majority, with a 'long tail' of less common classes. Like these studies, we find that it is often the 'long tail' of categories which contain the more interesting and controversial practices (for instance, the use of 'sensitive' data classes in the trading of personal data) which are commonly the focus of media attention and public concerns.

Our analysis appears to have revealed a number of uses of data which go some way to justify these public concerns. In particular, for 'trading / sharing in personal information', each of the ICO's eight kinds of sensitive data are apparently collected for this purpose by at least 200 organisations. Such practices are rare and might be unobjectionable if put into context. But concerned data subjects are unlikely to be reassured on the basis of the information provided in the register. In addition, the combination of certain purposes, data classes and recipients does in some cases imply a data processing arrangement which could be unlawful without a number of special compliance measures, and the register does not provide information on whether such measures are indeed in place.

In some cases, the legitimacy of such arrangements is questionable even if such measures were in place. For instance, it is difficult to imagine circumstances under which a credit reference agency would be justified in using an individual's political opinions as a basis for lending decisions (as 20 organisations in the register state), or why any data subject would knowingly consent to such a practice. While the appearance of such arrangements could be the result of mistakes made during the registration process, organisations would presumably wish to correct such information (or be prompted to do so by the ICO), given that it is publicly available. Another possibility in these situations is that the data may have been subjected to de-identification procedures prior to sharing it; in which case, data protection restrictions may not apply. However, the register does not detail whether such measures have been made. Furthermore, even if they have, concerned data subjects may nevertheless be concerned about their data being shared in supposedly 'anonymised' form, given the ease of re-identification in many cases (Narayanan & Shmatikov, 2007), (Ohm, 2010).

In contrast to some of the previous research which has shown significant differentiation between company privacy practices (Bonneau & Preibusch, 2010), (Cranor et al., 2013), we find a lack of variation in each of the three sectors we studied. For instance, where Cranor et al found just 24.4% of US financial institutions shared data with affiliates, we found that 93% of UK financial service providers did so[28]. While different regulatory environments

[28] In Cranor et al, affiliates are defined as entities 'related by common ownership or control' to the institution in question), while the register refers to 'Other companies in the same group as the data controller'.

and other conditions prevent any direct comparison between the UK and US banking sectors, this nevertheless indicates that for UK consumers who prefer financial service providers to not share their data with affiliates, significantly fewer appropriate choices appear to exist.

12. Limitations

Whilst it enables new analysis on an unprecedented scale, this data source is not without its limitations. A large portion of data processing occurs outside its scope; many companies whose data practices affect UK consumers, such as large international web companies, do not operate their consumer-facing services from their UK offices and therefore are not necessarily required to register (although this is a complex and changing area). In addition, processing of 'anonymous' data does not need to be disclosed in the register.

Another key issue is granularity. Many categories contained in the dataset would be more informative if given separate definitions. For instance, some consumers may perceive a difference between trading personal data for a profit and sharing it for some social purpose. As such, the standard description 'Trading / sharing personal information' is too broad.

Furthermore, the data within a purpose entry is not fine-grained enough. For instance, if the data subjects are 'customers' and 'staff', and the data classes are 'financial details' and 'physical and mental health', it matters greatly which of the data classes pertain to which data subject. It may be perfectly acceptable to collect certain data about staff but not customers, or vice-versa. As it stands, many of the entries in the register give the appearance of potentially unethical or illegal practices because of this lack of differentiation. This has been made worse by the new format, which, in an effort to make individual entries shorter and more user-friendly, has obscured which categories of data subjects, classes, and recipients are associated with which purposes, preventing any meaningful disclosure on a per-purpose basis.

Finally, one might be sceptical about the accuracy of some of the disclosures organisations make. In addition to the possibility that organisations may not reveal some of their practices in their notification, they may also have perverse incentives to claim engaging in additional practices that they do not really engage in. This is because there appear to be penalties for not disclosing practices that a controller is later found to be engaging in, but no penalties for listing a practice that a controller does not currently engage in but may at some point in the future. Therefore, a rational controller might be inclined to list as many categories as possible in order to cover themselves and avoid penalties for any activity they later take. A functioning market for privacy, where consumers actively seek out those organisations with better practices, would ideally mitigate this effect, by providing an incentive to organisations both to engage in more privacy-respecting practices, and at the same time report those practices accurately, including the limits of their practices.

13. Recommendations

With an increasing amount of personal data being collected and used, and with data-driven decision-making playing an increasingly important role in peoples lives, the need for the kind of transparency mechanisms studied in this paper is ever higher. Researching privacy notifications and the systems and standards that support them is therefore not only of academic interest but also has implications for policy and practice. As evidenced in the background section, a wide variety of technical work and policy proposals has pointed to the need for an appropriate standard for privacy-related disclosures.

With this in mind, a number of considerations for improving the design of future notification / transparency systems (some of which are outlined in the background section) arise from our analysis. As mentioned above, the register is now in a changed format, and is likely to be abandoned by the ICO as a result of impending changes to E.U. data protection law. Despite this, designers of future systems could benefit from understanding the features which make the ICO's register more or less useful.

First, the analysis herein clearly demonstrates the benefits of notifications being made available in a standard, machine-readable format. This significantly reduces the barriers to scaling up analysis of organisational privacy practices from individuals to whole sectors or countries. Previous research has had to rely on relatively inefficient methods, from writing special natural language parsing software for standard-form policies, or worse, manual analysis of full-length legal documents. Having machine-readable data to start with (even if it requires some additional parsing and processing), drastically reduces these barriers.

Second, division of privacy-related practices into the categories of purpose, subject, class, and recipient is useful, and all four categories seem necessary in order to derive any kind of meaningful conclusions about an organisations' practices. Any notification system which leaves one or more of these out is likely to prevent meaningful analysis. It is unsurprising, therefore, that other standards for such disclosures, such as P3P and the US financial institution model privacy form (Cranor et al., 2013), have included equivalent categories.

However, the provision of these categories alone is not necessarily sufficient. Most importantly, without further fine-grained differentiation, the notification is likely to leave significant ambiguity. Rather than lists of data subjects, classes, and recipients for one purpose, it would be far more informative to differentiate subjects, classes and recipients individually, rather than aggregating them on a per-purpose basis. This way, the notification would indicate exactly which classes apply to which subjects, and which recipients have access to which classes. This would increase the amount of input involved in each notification, and therefore be more onerous on the organisation making the disclosure. However, without such differentiation, a viewer of the notification is likely to be misled.

Changes made in April 2013 to the notification process for the UK register have unfortunately made the data contained within it even less fine-grained. Instead of differentiation on a per-purpose basis, distinct purposes and associated information about subjects, classes, and recipients, have been amalgamated into one entry. It is no longer possible to ascertain, for instance, the precise purpose or purposes that data about customers are gathered under, and if so, which categories of data are gathered for which purpose. So the problematic lack of granularity encountered in the data prior to April 2013 is now even greater, rendering the resource even less informative than it previously was.

Requiring per-field differentiation would also ideally go hand-in-hand with better incentives for accurate disclosures by organisations. As mentioned above, at present, organisations may assume (correctly or not) that they can reduce their legal liability by exaggerating the extent of their actual practices. There are numerous ways organisations might be encouraged to make more accurate and detailed disclosures, from improved guidance for registration, to mandatory audits. However, one measure would be for regulators to pro-actively monitor the content of each notification using the kinds of techniques explored here, using this as a basis for further investigation.

Previous research has noted the opportunity this kind of analysis presents for improving regulatory practice. Having found evidence of contradictory and potentially illegal practices in the disclosures of financial institutions, one group of researchers suggest that failure to identify and act on such evidence is a missed opportunity on the part of the regulator (Cranor et al., 2013). They ask; 'if we as academics can quickly uncover these issues, why have regulators who are charged with overseeing these financial institutions not already done so?'. The same could be asked of the ICO. For instance, discovering any organisation who, in their registration, claims to use sensitive personal data to make credit reference decisions (as was the case for over a hundred organisations in this study) could ring the regulator's alarm bells. This information could generate an automated message to the data controller to ask whether appropriate special measures have been put in place to ensure the processing is legitimate. This kind of targeted action would be onerous if it involved manually checking every one of the over 350,000 registrations for potential non-compliance, and may be beyond the capacity of even a well-resourced regulator. But as demonstrated here, machine-readability means that by employing simple analytical techniques on their own data, regulators or other entities could automatically identify and contact potentially non-compliant notifications.

In addition to pressure from regulators, informed decision-making by privacy-conscious consumers is also likely to pressure organisations to make more accurate notifications, and develop more privacy-friendly practices. At present, this is prevented by a lack of transparency and consumer ignorance of organisation's practices. Those consumers who are concerned about their privacy do not have the means to make informed and meaningful choices between service providers. While a better notification system may not in itself change this, it could provide the basis for intermediary services which would rate organisations practices on behalf of privacy-conscious consumers, and in turn provide a commercial incentive for organisations to improve practices.

14. Conclusions

Transparency is easy to affirm but hard to achieve in practice. There has been no shortage of enthusiasm for measures which render visible organisations' policies and practices regarding personal data. We are left with a graveyard of incomplete attempts. The data source studied here is arguably the largest and most complete arising from any of them, and therefore provides an instructive case study through which we can assess the viability of this transparency project. The analysis above demonstrates some of its potential. This kind of resource does enable macro-level conclusions about the types of data used by a range of organisations, the purposes involved, and the types of recipients with access to the data. This information could be an important starting point for more detailed investigations.

However, the data source itself is not designed for a nuanced understanding at the level of particular organisations' practices. There are two kinds of problems associated with using these broad, abstract descriptions as a means to detect particular practices; false negatives, where the data fails to capture the existence of a practice, and false positives, where the data suggests a certain practice is occurring where it is not. The ultimate utility of this resource may therefore depend on whether there is value in this macro-level abstraction despite the strong possibility of errors. When it comes to describing the use of personal data, there will always be tension between standardisation and nuance, abstraction and detail. Resolving these tensions may be the key to successful transparency systems in this domain.

References

Beatty, P., Reay, I., Dick, S., & Miller, J. (2007). P3P Adoption on E-Commerce Web sites, (April), 65–71.

Berners-lee, T. (2004). Semantic Web Road map Machine-Understandable information : Semantic, (September 1998), 1–10.

Bonneau, J., & Preibusch, S. (2010). The Privacy Jungle: On the Market for Data Protection in Social Networks. Economics of Information Security and Privacy, 121–167. doi:10.1007/978-1-4419-6967-5_8

Bradwell, P. (2010). Private Lives: A People's Enquiry into Personal Information. Retrieved from http://www.demos.co.uk/publications/privatelives

Brockdorff, N., & Appleby-arnold, S. (2011). CONSENT: What Consumers think.

Byers, S., Cranor, L. F., Kormann, D., Ave, P., & Park, F. (n.d.). Automated Analysis of P3P-Enabled Web Sites.

Cranor, L. F. (2012). NECESSARY BUT NOT SUFFICIENT : STANDARDIZED MECHANISMS FOR PRIVACY NOTICE AND CHOICE, 273–307.

Cranor, L. F. (2013). P3P Original Idea behind P3P.

Cranor, L. F., Egelman, S., & Sheng, S. (2008). P3P Deployment on Websites P3P Deployment on Websites.

Cranor, L. F., Idouchi, K., Leon, P. G., Sleeper, M., & Ur, B. (2013). Are They Actually Any Different ? Comparing Thousands of Financial Institutions ' Privacy Practices. The Twelfth Workshop on the Economics of Information Se-Curity (WEIS 2013) , June 11–12, 2013, Washington, DC.

Culnan, M. J. (2000). Protecting Privacy Online: Is Self-Regulation Working? Journal of Public Policy & Marketing, 19(1), 20–26. doi:10.1509/jppm.19.1.20.16944

Egelman, S., Tsai, J., Acquisti, A., & Cranor, L. F. (n.d.). Studying the Impact of Privacy Information on Online Purchase Decisions, 1–4.

Federal Trade Commission. (2013). FTC Testifies on Data Brokers Before Senate Committee on Commerce, Science and Transportation. Retrieved from http://www.ftc.gov/news-events/press-releases/2013/12/ftc-testifies-data-brokers-senate-committee-commerce-science

Gandon, F. L. (2003). Semantic web technologies to reconcile privacy and context awareness.

Gutwirth, S., & DeHert, P. (2006). Privacy, Data Protection and Law

Enforcement. Opacity of the Individual and Transparency of Power.

Greenleaf, G. (2012). Global Data Privacy in A Networked World, in Brown, I. (ed) Research Handbook on Governance of the Internet Cheltenham: Edward Elgar,

Kelley, P. G., Cesca, L., Bresee, J., & Cranor, L. F. (2009). Standardizing Privacy Notices : An Online Study of the Nutrition Label Approach Standardizing Privacy Notices : An Online Study of the Nutrition Label Approach.

Leon, P. G., Cranor, L. F., Mcdonald, A. M., & Mcguire, R. (2010). Token Attempt : The Misrepresentation of Website Privacy Policies through the Misuse of P3P Compact Policy Tokens Token Attempt : The Misrepresentation of Website Privacy Policies through the Misuse of P3P Compact Policy Tokens.

McDonald, A., & Cranor, L. (2008). Cost of Reading Privacy Policies, The. ISJLP, 0389, 1–22. Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/isjlpsoc4&section=27

Milne, G. R., & Culnan, M. J. (2002). Using the Content of Online Privacy Notices to Inform Public Policy: A Longitudinal Analysis of the 1998-2001 U.S. Web Surveys. The Information Society, 18(5), 345–359. doi:10.1080/01972240290108168

Narayanan, A., & Shmatikov, V. (2007). Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset).

National Telecommunications and Information Administration (NTIA). (2013). SHORT FORM NOTICE CODE OF CONDUCT TO PROMOTE TRANSPARENCY.

OECD. OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980). Retrieved from

http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review.

Schwartz, A. (2009). Looking back at P3P: lessons for the future, (November). Retrieved from https://www.cdt.info/files/pdfs/P3P_Retro_Final_0.pdf

Tsai, J. Y., Egelman, S., Cranor, L., & Acquisti, A. (2010). The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study. Information Systems Research, 22(2), 254–268. doi:10.1287/isre.1090.0260

U.S. White House Office of Management and Budget. (1983). The President's Annual Report on the Agencies' Implementation of the Privacy Act of 1974 (p. at 118 (Dec. 4, 1985).).

United States Government Accountability Office. (2013). INFORMATION RESELLERS Consumer Privacy Framework Needs to Reflect Changes in Technology and the Marketplace Consumer Privacy Framework Needs to Reflect, (September).

Vimercati, G., Paraboschi, S., & Pedrini, E. (2009). Primelife policy language. Retrieved from http://spdp.di.unimi.it/papers/w3c_wsacas_2009_02.pdf

Wellcome Trust. (2013). Summary Report of Qualitative Research into Public Attitudes to Personal Data and Linking Personal Data. Retrieved from http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp053205.pdf