Article

JIS

Journal of Information Science
2018, Vol. 44(4) 552–565

DOI: 10.1177/0165551517706233
journals.sagepub.com/home/jis
SAGE

The digitally extended self: A lexicological analysis of personal data

Brian Parkinson
School of Electronics and Computer Science, University of Southampton, UK

David E Millard
School of Electronics and Computer Science, University of Southampton, UK

Kieron O'Hara
School of Electronics and Computer Sciences, University of Southampton, UK

Richard Giordano
Faculty of Health Science, University of Southampton, UK

Abstract
Individual's privacy, especially with regard to their personal data, is increasingly an area of concern as people interact with a wider and more pervasive set of digital services. Unfortunately, the terminology around personal data is used inconsistently, the concepts are unclear and there is a poor understanding of their relationships. This is a challenge to those who need to discuss personal data in precise terms, for example, legislators, academics and service providers who seek informed consent from their users. In this article, we present a lexicological analysis of the terms used to describe personal data, use this analysis to identify common concepts and propose a model of the digitally extended self that shows how these concepts of personal data fit together. We then validate the model against key publications and show in practice how it can be used to describe personal data in three scenarios. Our work shows that there is no clearly delineated kernel of personal data, but rather that there are layers of personal data, with different qualities, sources and claims of ownership, which extend out from the individual and form the digitally extended self.

Keywords
Categorisation; Data Protection Act; digital footprint; digital mosaic; digital persona; digitally extended self; informed consent; personal data; privacy; virtual self

1. Introduction

The continuing concerns about the use of personal data, especially with respect to privacy, informed consent and right of access to data, drive a need for well-defined and consistent terms to describe those data. This article focuses on data that are descriptive of an individual. The increasing use of this type of personal data, and the resulting markets in personal data [1], has led to concerns regarding issues of privacy [2]; privacy-related decision-making [3]; informed consent for organisations to collect, process, curate and transfer their data to other bodies [4]; and also an individual's right of access to data descriptive of them [5].

Given these concerns, it is surprising that there is no common terminology around personal data. What nomenclature should be used for digital data that is descriptive of an individual? What collective nouns can be used to classify the data and how are they related to each other? A variety of terms present themselves in the literature, for example, digital

Corresponding author:
Brian Parkinson, Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK.
Email: BLP1M11@soton.ac.uk

footstep, fingerprint, *shadow*, *profile*, *mosaic*, *persona*, *virtual self* or *doppelgänger*. The terms are widely used but not in a consistent way. Neither are the usages critiqued. The problem of a common set of terms in the face of technological change has been noted before; for example, Bakshi [6] highlights inconsistencies in use when discussing the digital economy in general, and Heinderyckx [7] points to rapid rate of change of information and communication technology (ICT) terminology. Others have tried to address the problem in other domains; for example, Safran et al. [8] defined terms when discussing health data, but none has aimed to specifically discuss the terminology associated with personal information.

The use of consistent terms and concepts is important because it reduces ambiguity in academic debate and improves information sharing – particularly between service providers and their users. When giving informed consent, an individual must determine, and understand, the information that is covered by the agreement [9]. It is also important for legislators to evaluate and use terminologies consistently, whether they be incorporated in organisational privacy agreements or legislative language or guidelines. In addition, a concrete set of concepts is important for the design of systems that deal with personal data, as it may have implications for how data are stored, managed and exposed through a range of interfaces.

In order to tackle this problem, we set out to analyse the terminology and concepts of personal data present in the literature, with the goal of identifying common concepts (even when they are named differently) and establishing their relationships. We then selected the most popular/descriptive terms and bring these concepts together in a model of the digitally extended self. The model is then tested against literature and against data descriptive of the first author. This illustrates two uses, the first as a standard set of terms and the second as a high-level data model.

The remainder of the article is arranged as follows: Section 2 describes the related work in data classification and modelling, Section 3 describes the method used for our lexicological analysis of the literature, Section 4 discusses the terms encountered and their relationships and Section 5 shows how these can be brought together in a coherent model. Section 6 presents a validation of the model against 45 key publications from the original sample and shows how the model can be applied to a particular scenario. Section 7 concludes the article and draws out implications for future developments.

2. Related work

The personal, or social, point of view is generally used when framing the debate regarding issues of privacy and data descriptive of an individual [10]. However, other perspectives may be adopted. For instance, Pollach [11] suggests a function-based approach, forming a matrix of data types (e.g. sales data) and data handling methods (e.g. selling) in order to help people better understand the consent that they are giving. However, as a method of classification for all data descriptive of an individual, we find this approach limiting due to the constraints of constructing an exhaustive set of types and processes. A similar approach, used in the UK Data Protection Act 1998, considers not only types of organisation that hold data (e.g. research organisations) but also the use to which the data are put (e.g. domestic purposes). Again, this approach does not provide an exhaustive classification of data descriptive of an individual, and it can be argued that the data covered are in parts unclear [12]. This may be a cause of inconsistencies in the categories of data provided that are found in responses by companies to subject access requests under the Act.

Polonetsky et al. [13] take a more ontological approach and propose a categorisation of personal data based upon degrees of identifiability of an individual. This is a useful contribution to the vexed problem of big data usage and personal privacy, and the approach does provide a complete classification. However, what may appear to be de-identified data today may be identifiable tomorrow due to technical advances such as the use of additional data sets that compromise the level of anonymity of the data. Consequently, the classification of data based on degrees of identifiability may fluctuate and become indeterminate.

An alternate approach to data descriptive of an individual would be through the data information knowledge wisdom (DIKW) hierarchy, a structural approach to data and its transformational uses. The assumption is that data at the bottom of the hierarchy are transformed through processing into information, which is processed to create knowledge, and knowledge, in turn, yields wisdom [14]. This structural and transformational framing can be used to argue that data by itself offer no threat to privacy unless they can be transformed into information, knowledge or wisdom, each having the potential to be more threatening to an individual's privacy. While Batra [15] argues that the advent of data analytics in real time blurs the DIKW distinctions, the classification is still of some interest as not all data are subject to analytics, and those that are can still be classified.

Finally, Palfrey and Gasser [16] use availability as a classification tool, observing a distinction between data that are publicly available and those that are not. This is used to differentiate between the digital identity (the publicly available) and digital dossier (all data descriptive of an individual). There are two issues with this classification: it may be

considered too simplistic a distinction if it were the only observation made, and more significantly, it does not have clear boundaries. For example, that which is considered available by a computer-literate person would be different to that accessible to others with more limited skills.

In our work, we have taken a new approach that is based upon the origin, handling and manipulation of data by various agents associated with an individual. We will demonstrate that this has the benefit of communicating ways in which personal data are transformed and transported while providing a full categorisation of the domain and at the same time being readily understandable.

3. Method

An initial search of the privacy and surveillance literature enabled us to extract a list of terms used to describe facets of data descriptive of an individual. In order to perform a lexical analysis of the meanings allocated to these terms, it was necessary to obtain examples of their usage. Several data sources for the search were considered, for example, Web of Science, Scopus or university-specific search engines such as Oxford's SOLO. Google Scholar was selected due to the wide range of papers and books within its base of data, the ease of integration into the chosen reference manager (Bookends) and its increasing use within the research community [17,18]. Its weakness with respect to Boolean searches and the restriction to 1000 search results [19] were not significant for this research.

In total, we identified a set of four common starting terms from our initial literature review (*digital footprint*, *digital mosaic*, *digital persona* and *virtual self*), and these then became our seed search terms for Google Scholar. The search engine, at the time of this work, returned a maximum of 1000 items for each search, and so for high usage terms we searched by calendar year, thus maximising the number of references returned. For each term, we then ordered the results by citation (discovering a power law distribution, meaning that each term had a relatively small number of higher cited sources). We then selected a purposeful sample from these based on high citations relative to publication date and overall size of the sample for that term. Terms and their meanings were then manually extracted and analysed. Where new terms were discovered, they were added to the list to be researched and the process undertaken again, resulting in a snowball sample of 64,584 papers covering 16 search terms and resulting in a purposive sample of 247 (the terms are shown in Table 1 together with the total count of results and the number of papers selected under each term for the purposive sample).

Digital fingerprint and second self have a relatively low purposive sample due to the high number of spurious results. For instance, digital fingerprint is a common term within forensic science, and second self is part of common phrases such as 'the second self-control task' and 'Barber's second self-creation theory'.

In order to determine the usage, each term was taken in turn, and the sample documents, containing that term, were examined. Meanings were observed and common themes were extracted. The terminology descriptive of personal data was then examined, and through a series of iterations, involving the second author, a standard categorisation was developed from which the overarching data model was derived. The naming of these categories was based upon common usage and strength of metaphor. A further iteration to validate the findings was then undertaken and is described in Section 6.

A potential weakness of this approach is the dependence upon the work of the first author to examine the literature and extract meaning. It can be argued that the use of a second researcher to independently analyse the literature and identify themes would strengthen the findings. However, as Armstrong et al. [20] note, this type of analysis is a form of interpretation in which researcher's views have important effects. It is possible that a second researcher may have come to a different, but no more valid, conclusion. The derivation of categories and their labels was, however, subject to iterative

Table 1. Summary of Google Scholar search results, August 2014.

Search term	Google Scholar	Purposive sample	Search term	Google Scholar	Purposive sample
Digital biography	165	4	Digital persona	1679	18
Digital doppelganger	77	11	Digital self	4223	23
Digital dossier	421	4	Digital shadow	592	4
Digital fingerprints	4930	2	Ersatz double	11	1
Digital footprint	1501	29	Online identity	9256	18
Digital identity	9059	26	Second self	25,578	12
Digital mosaic	834	11	Shadow identity	171	2
Digital person	967	39	Virtual self	5120	43

debate between the authors with the objective of producing a consistent set of terms that can be used when discussing personal data. While others may have decided on an alternate nomenclature, we have endeavoured to create a categorisation and set of names that are informative, easy to understand and remember [21].

4. Results

The analysis of the terminology and their usage identified three main issues. First, terms used to describe categories of data descriptive of an individual are also used to label other things; second, single noun phrases were used to label similar but differing groupings; and third, more than one noun phrase was used to label a single grouping.

4.1. Terms used to describe categories of data also used to label other things

It must be expected that variations in usage will be identified when examining the use of 16 noun phrases, and within a discourse, the meanings tend to cover overlapping sets of things, as we will present below. However, when analysing the usage across discourses, as we did in this case, then examples of entirely different meaning were observed. Digital footprint, for example, is used to label the outline of a building on a digital map [22]. Digital mosaic may be a collection of images used to create a larger image as in the case when illustrating the location of dengue fever in Nicaragua [23] or a collection of videos, which together form a composite video [24]. In another discourse, virtual self was used by Goffman [25] to describe a role acted by an individual in their everyday life, while Metzinger [26] used the term to cover phantom limbs, dream states and out-of-body experiences and Valk [27] considered that there are no humans in the world but that we all exist only in an immersive virtual environment as virtual selves. When terms are used across disciplines to label separate things, as we have illustrated above, meaning is created through use and explanation. However, when a single term is used to categorise similar but differing things, it becomes imprecise and hence problematic.

4.2. Terms with multiple meanings

When looking at meaning in the discourses surrounding data that are descriptive of an individual, variations in meaning were observed. Rather than exhaustively listing these, the following illustrative examples are presented.

The term *digital footprint* is used to categorise data left behind by an individual in the virtual world [28,29]. The emphasis is on an individual leaving their own data trails. However, Palfrey and Gasser [16], among others, state that

> [d]igital footprints are digital artifacts which can be left by the individual or by another. (p. 33)

Sellen et al. [30], however, state that digital footprints are created about which the individual has little or no knowledge or control. This raises the question of whether the subject individual, another individual or both create digital footprints and whether the subject individual knows of them or not.

A second noun phrase in common use is *virtual self*, which Lyon [31] uses to identify collections of data, and analyses, that describe an individual and which Turkle [32] sees as

> extensions of ourselves we have embodied in program. (p. 166)

There may be a single virtual self to represent all data and analyses descriptive of an individual, or else, multiple virtual selves representing subdivisions of the data and thus perhaps replicating Goffman's contextual self-projections within the 'real world' [33]. However, is there is a difference between the actors creating the virtual self or of an organisation imposing a persona upon an individual? Lyon would appear to consider the virtual self as imposed perhaps as a result of some form of surveillance or analytics. Turkle [34] considers the individual as the creator or persona(s). Indeed, this is the case with Pearson [35] who describes the virtual self as a constructed online identity, while Bessière et al. [36] use the term to label the self created within an online game and give it the synonym of *avatar*. The virtual self can, however, also be distinguished in another way, either as representing a para-authentic extension of the individual or as a construction of an alternate personality [37], perhaps as an experimental device. Finally, a less complex projection of the self is a photograph used to represent an individual, for example, on a social network site, but labelled a virtual self [38].

4.3. Multiple terms same meaning

We have provided two examples of terms used to describe similar but differing categories of data. There is, however, a situation where multiple terms are used to describe the same thing. In the case of categorisation of data descriptive of individuals, the use of different terms to identify the same class of elements can cause uncertainty and a resultant lack of rigour. This is demonstrated in Section 6.1. For instance, *digital footprint* [39], *digital fingerprints* [40] and *digital persona* [41], as used in the cited papers, are all synonyms and used within the context of personal data. In this instance, the use of the more commonly found term *digital footprint* would provide consistency and allow the nuanced inference of an individual's digital artefacts being used to create an online persona to be explained more fully.

4.4. Summary

We have illustrated a lack of consistency in the application of noun phrases used to label categories of data descriptive of an individual. One way forward would be to leave the situation unchanged and allow usage to either continue in an unclear way and hope that time will allow meanings to coalesce around the most popular noun phrases while others wither and die away. We have taken an alternate approach and have developed a classification model for the data descriptive of an individual, which we present below.

The noun phrases chosen to label categories of data were selected as a result of their commonality of use and strength of metaphor. For instance, *digital footprint* and *digital fingerprint* both describe a data artefact left by some activity of an individual, which reflects itself in the virtual world. Footprints in the sand tend to disappear and can be readily observed, although they cannot normally identify an individual. On the other hand, fingerprints tend to remain for many years, can identify an individual and are difficult to observe. In this case, although *digital fingerprint* is the stronger metaphor, *digital footprint* was selected because it is used more widely.

The terms selected were then developed into a coherent and consistent categorisation of an individual's data as it is represented in the virtual world. These are described in the following section and illustrate the gradation of personal data as it is deposited, merged, transformed and analysed.

5. Model

In our total sample, we identified 16 terms, but in our analysis, we were able to group these terms into one of five categories, each of which we argue is distinctive in terms of its origin and construction and which together form the layers of our model. We named each category after the term that we felt was most representative. The five concepts in our model are as follows:

- *Digital footprint*: data descriptive of an individual, laid down by that individual as a result of using, or being observed by, computing devices;
- *Third-party digital footprint*: digital footprints created by an individual, or a computer system, which are descriptive of another individual (the data subject);
- *Digital mosaic*: a collection of digital footprints which can be used to create a picture of a person; a simple digital mosaic consists of a person's own digital footprints, whereas a full digital mosaic is used to describe the collection of both an individual's own and third-party digital footprints;
- *Digital persona*: a model of an individual created by the analysis of digital footprints and/or other digital personas and optionally additional second-level data;
- *Digitally extended self*: the combination of the above elements to provide the fullest possible digital representation of an individual.

In our definitions, the term *second-level data* is used to identify data that are not directly descriptive of an individual but which provide information about an individual's attributes (e.g. demographic data which are associated with a person's post/ZIP code).

There are several reasons for placing these categories together in a coherent model. First, it provides a vehicle for discussing the issues associated with the collection and use of an individual's data, and in doing so defines a set of terms, thus reducing ambiguity. Second, it illustrates where boundaries exist. It is often at the edges where more interesting and difficult decisions have to be made, especially with respect to knowledge and control of an individual's data. Finally, by naming structures in certain ways, how they are viewed is affected. In this case, the term 'digitally extended self' has been created to describe the virtual self, not as a separate entity but as an extension of the real self.

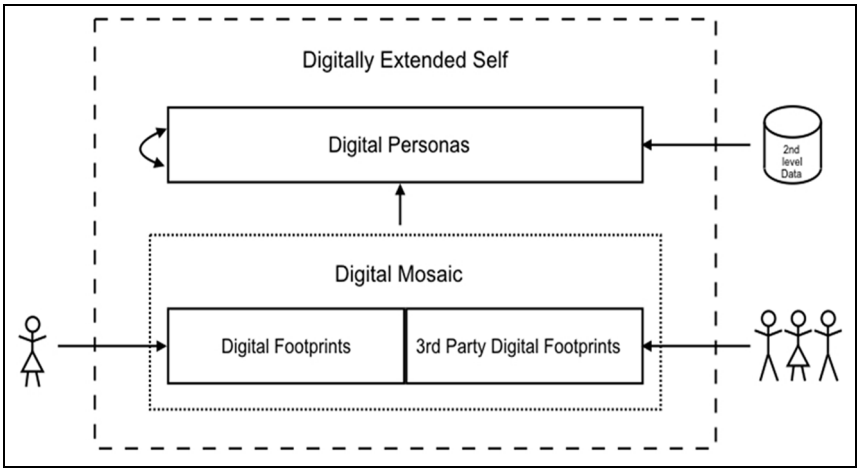

Figure 1. Our model of the digitally extended self – showing the five categories of personal data.

Figure 1 shows an overview of the model. The basis of the model are the digital footprints created by the data subject and the third-party digital footprints created by other individuals. Combined they form the digital mosaic. This, in turn, is the basis for digital personas that typically exist to profile an individual for some purpose. These personas may also use second-level data and other digital personas as input to the analysis. The whole is then defined as the *digitally extended self*.

6. Validation of the model

The model serves two purposes. It provides a clear nomenclature which facilitates a cross-disciplinary use of terms, and the second is as an overarching data model. The model is therefore validated in two ways: first, to ensure that the model encompasses the existing, highly variable and disorganised terminology, and second, against actual data.

6.1. Validation against terminology

As a first validation step, we show (Tables 2–6) how a range of terms and usages from the purposive sample map to the categories in our model. To create this mapping, we selected 45 examples that provided coverage of the model concepts and where the same terms are used in different senses (e.g. Byron [28] discusses digital footprints in the same way as our model, but Chretien et al. [43] use the term to describe something that maps to a digital mosaic in the model instead). Within the publications, no match for third-party digital footprint was found as the phenomena were mentioned but not named; it is therefore omitted from Tables 2–6.

With this exception, it was possible to exhaustively map terms found in the literature sample to the categories proposed as a result of the analysis, showing that all the terms used in the 45 publications, that refer to an individual's data, map to specific parts of the proposed model.

The validation above shows that all the categories within the model map to phenomena that have been discussed in the literature and gives a sense of the ways in which different terms have been used to express and describe them. The second scenario-based validation comes to the model from the other direction and looks at how the data involved in real case studies map to the model. It thus demonstrates that the model covers all of the data in the case studies and also shows how the distinctions made by the model are useful for discussing data in that particular scenario.

6.2. Validation against actual data

The case studies we have used are based on the interactions of the first author with a UK-based bank, an international charity and a credit reference company. To create the case studies, data were gathered through subject access requests[1] made to the organisation in September 2013, March 2014 and October 2014. This case study is part of a wider data collection exercise in digital autoethnography during which the first author requested data, under UK and European law, from 32 organisations selected from over 400 with which he had interacted. The organisations were chosen to represent

Table 2. Digital footprint: mapping of literature to the model.

Term in model	Term from literature	Usage	Example of usage
2 Digital footprint	2.1 Digital fingerprints	2.1.1 'Data about individuals held in the hands of third parties' (p. 2)	Wittes [40]
	2.2 Digital footprint	2.2.1 A digital artefact left behind by some activity 'as they "tread" through the World Wide Web, they leave behind a "footprint"' (p. 1227)	Batchelor et al. [39] Siemens and Long [42] Greysen et al. [29]
		2.2.2 'Personal information available online' (p. 58)	Byron [28]
		2.2.3 Postings on social media (by medical students)	Chretien et al. [43]
		2.2.4 Patterns of Internet usage/artefacts	Hankin [44]
		2.2.5 Traces of online presence	Hengstler [45] Madden et al. [46] O'Keeffe and Clarke-Pearson [47]
		2.2.6 Pervasive environments and contextual traces	Kapadia et al. [48]
		2.2.7 Results of activity in the virtual world which describes someone	Palfrey and Gasser [16]
		2.2.8 A group of digital footprints on one site, i.e., Facebook	Moore and McElroy [49]
	2.3 Digital persona	2.3.1 An electronic portfolio of work created by a student	Clark [41]
	2.4 Identity	2.4.1 'A trail of data artifacts' (p. 10)	Briggs [50]

Table 3. Simple digital mosaic: mapping of literature to the model.

Term in model	Term from literature	Usage	Example of usage
3 Simple digital mosaic	3.1 Digital dossier	3.1.1 Dossiers compiled from a person's uploads	Gelman [51]
	3.2 Digital footprint	3.3.1 Referring to the collection of digital footprints	Chretien et al. [43] Ess [52]
	3.3 Digital mosaic	3.2.1 'He was a digital mosaic … storing his data in starfish satellites' (p. 112)	DeLillo [53]
		3.2.2 Google search terms used by an individual and their associated data	Floridi [54]
		3.2.3 'Our transactions, our media consumption, our locations and travel, our communications, and our relationships' (p. 2)	Wittes [40]

central and local government, public and private companies, and non-governmental organisations (NGOs) across a range of sectors (e.g. credit reference, online retail and utilities). The first author requested all information held that was descriptive of him, including analyses, where data were obtained and where they were sent. A snowball sample was then created from responses that included details of bodies providing data to, or receiving data from, organisations in the original purposive sample. The snowballing process terminated on a pre-defined cut-off date by which time requests had been sent to a total of 82 organisations. These analyses were then compared with the model described above. Follow-up requests were then made asking for data that had been omitted. The cumulative results of the requests were organised in terms of our model. Altogether, 82 models were successfully created. We then selected three of these organisations as our case studies, based on their ability to illustrate the different aspects of the model.

Figures 2–4 show the diagrams of what these data were in our case study examples and where they fit within the model. In this instance, we use a centric diagram, to illustrate *digital footprints* at the heart of the data, which describes an individual. It is then incrementally extended through the concept of multiple artefacts forming a *digital mosaic*, identified by the inner circle. Next analyses are formed using data from digital footprints and external sources resulting in *digital personas*. The whole within the outer circle is named as the *digital extended self*.

6.2.1. Case study 1, a UK-based bank. The case study shown in Figure 2 takes data provided by a UK-based bank and maps it against our centric model of the digitally extended self, thus revealing the significance of the parts of the

Table 4. Full digital mosaic: mapping of literature to the model.

Term in model	Term from literature	Usage	Example of usage
4 Full digital mosaic	4.1 Data shadow	4.1.1 Combination of digital artefacts	Westin and Ruebhausen [55] Garfinkel [56] Floridi [57]
		4.1.2 'Records and data about the self' (p. 167)	Smithson [58]
	4.2 Digital biography	4.2.1 'An electronic collage' (p. 1394), 'a life captured in records' (p. 1394), 'bits and pieces of stored information about one's life' (p. 70)	Solove [59] Ploeg [60]
	4.3 Digital doppelganger	4.3.1 A collection of digital artefacts which provide a picture of a life	Cherry [61]
	4.4 Digital dossier	4.4.1 Collections of footprints e.g. from Facebook	Gross and Acquisti [62] Garfinkel [56] Solove [63]
	4.5 Digital footprint	4.5.1 Digital artefacts some known to us others not	Sellen et al. [30]
	4.6 Digital identity	4.6.1 Aggregated data about an individual, but only that which are publicly available	Palfrey and Gasser [16]
	4.7 Digital mosaic	4.7.1 Individual searches by law enforcement agencies may not intrude an individual's privacy, but multiple ones produce a mosaic of information which can be a breach of privacy	Dennis [64]
		4.7.2 A collection of artefacts which can present an image of an artist to a fan, e.g., YouTube, Twitter	Hanna et al. [65]
		4.7.3 Mosaic of information and analyses that create a picture of a company	Schwartau [66]
	4.8 Digital person	4.8.1 'A life captured in records' (p. 1)	Solove [63]
	4.9 Digital persona	4.9.1 Persona created by postings onto the Internet – does not consider analyses of these postings	Clark [41]
		4.9.2 'Each digital persona is defined by the combination of profile data captured by the person and others into one or more SNS [social networking systems]' (p. 11)	Clarke [67]
	4.10 Dossier	4.10.1 Aggregated data about an individual	Solove [68]
		4.10.2 Aggregated data about individual, includes data not publicly available	Palfrey and Gasser [16]

Table 5. Digital persona: mapping of literature to the model.

Term in model	Term from literature	Usage	Example of usage
5 Digital persona	5.1 Digital biography	5.1.1 'Bits and pieces of stored information about my life and behavior, an embodied identity' (p. 70)	Ploeg [60]
		5.1.2 Data and profiles	Solove [63]
	5.2 Digital – doppelganger, digital self, second self	5.2.1 Focuses on data from social networks and data aggregation	Andrews [69]
	5.3 Digital persona	5.3.1 Describes projected and imposed personae but does not explicitly include profile data, but does consider context	Ardagna et al. [70]
		5.3.2 Analysis of data, especially transaction-generated data	Blanchette and Johnson [71]
		5.3.3 Personas derived from profiling and data mining	Hildebrandt and Gutwirth [72]
	5.4 Digital persona, data shadow, digital individual	5.4.1 'The digital persona is a model of the individual established through the collection, storage and analysis of data about that person' (p. 1)	Clarke [73]
	5.5 Digital personality profile	5.5.1 'Aggregating, analyzing, or "mining" personal information, when it is or can be used to uniquely identify, locate, or contact that person' (p. 142)	Ludington [74]
	5.6 Ersatz double	5.6.1 Facebook profiles and postings	Sanchez [75]
	5.7 Online identity, digital self	5.7.1 Does not explicitly allow for the inclusion of profile data in further profiles	Briggs [50]

Table 6. Digitally extended self: mapping of literature to the model.

Term in model	Term from literature	Usage	Example of usage
6 Digitally extended self	6.1 Digital doppelganger	6.1.1 A similar concept but constrained to social networking data	Andrews [69]
	6.2 Digital dossier	6.2.1 'Taken together, all the digital information held, in many different hands, about a given person' (p. 39)	Palfrey and Gasser [16]
	6.3 Digital persona	6.3.1 Analysis of transactional data combined with other records, e.g., demographics	Blanchette and Johnson [71]
	6.4 Virtual self	6.4.1 Analyses computed by marketing companies and government departments augmented by further transactions	Lyon [31]

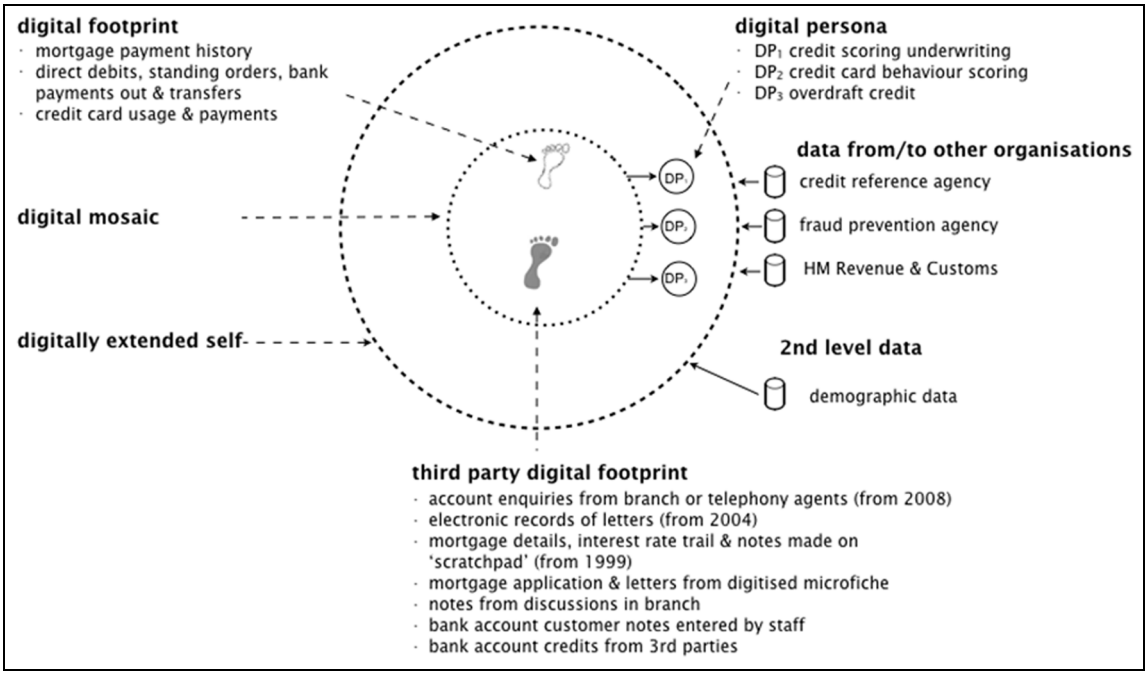

Figure 2. The centric diagram showing data (as instances of the model) from case study 1, a UK-based bank.

digitally extended self *not* under the direct control of the user – in this case, extensive notes and internal records of non-digital interactions made by third-party individuals (e.g. account enquiries from branch or telephony agents and bank account customer notes); three separate personas generated for purposes of underwriting, credit scoring and overdraft scoring; data that are independent of the individual and provided by third-party credit reference, fraud and taxation organisations; and, finally, second-level demographic data which describe the individual by inference to the location of their home.

6.2.2. Case study 2, an international charity. The second case study (Figure 3) uses data provided by an international campaigning charity and illustrates that although a minimal level of data was held by this organisation (name, email address and date of donation), a digital persona, received from an external company, was kept, showing propensities to open and to click on emails from this charity. In addition, the diagram illustrates that data are sent to the charity's offices in three other countries, two of which are approved by the European Union (EU) for the flow of personal data and one that is not, raising possible privacy concerns.

6.2.3. Case study 3, a credit reference company. The final case study (Figure 4) is a credit referencing organisation. This private limited company had no direct contact with the data subject but collected third-party digital footprints in the

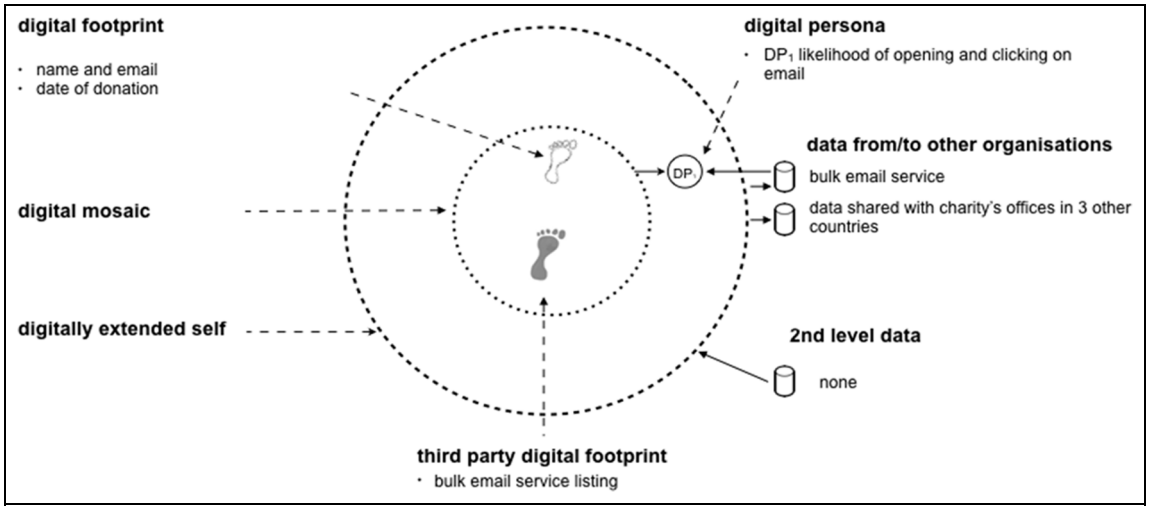

Figure 3. Case study 2, an international charity.

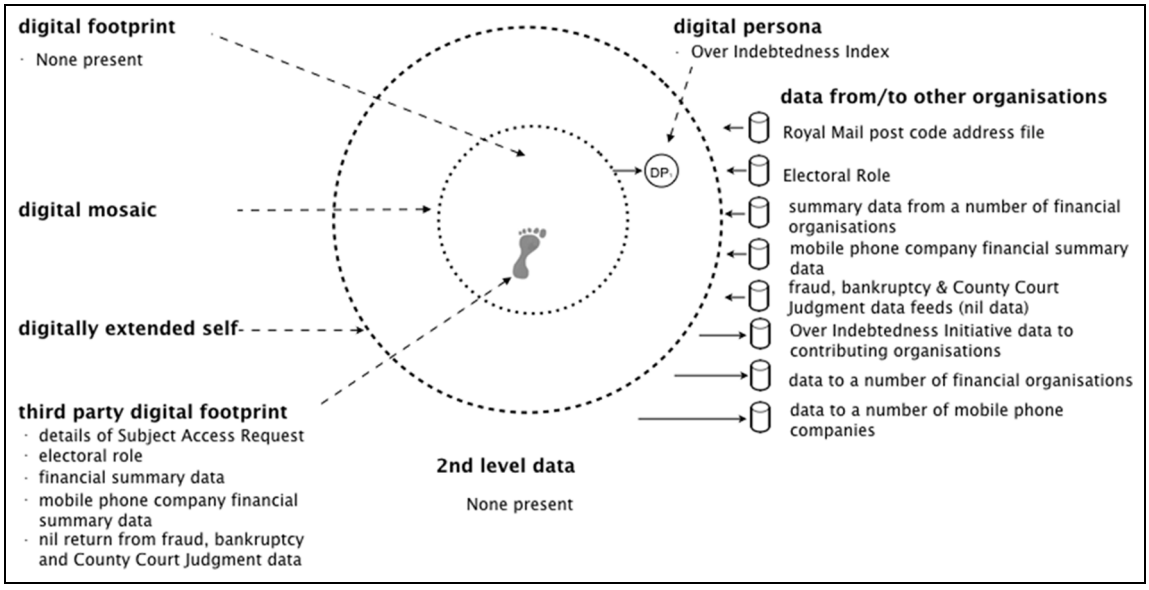

Figure 4. Case study 3, a credit reference company.

form of summary data from financial and mobile phone companies, which it combined with post code, electoral role, fraud, bankruptcy and court judgement data. This information was used to calculate an over indebted assessment which, together with other data, was provided to financial institutions and mobile phone companies. This collection and dispersal of data descriptive of the first author illustrates how sharing of personal data can be used to create a profile which is in turn distributed to other actors, unknown to the subject individual. This case also illustrates how the absence, rather than the presence, of a third-party digital footprint can itself be descriptive of an individual. The absence of fraud data, bankruptcy or county court judgements supports a mosaic (which fortunately showed the first author in a positive light).

6.2.4. Summary. Third-party footprints were not explicitly discussed in the literature (our first validation), but here they are shown to be an essential and extensive part of the description; indeed, in our third case study, they are the only components of the digital mosaic. These case studies illustrate how the model not only distinguishes between different types of data but also helps draw attention to the fact that an individual's digitally extended self is not tightly controlled or

atomic, but rather exists in graduated layers, with multiple owners, that progressively become less direct and more speculative as the data become more distant from the individual. In this context of multiple actors and varying gradations of data, which may be considered personal, it is clear why questions of privacy, ownership and rights of access are so complex.

To illustrate this, in the United Kingdom, there has been an on-going debate regarding the definition of personal data. The 1998 Data Protection Act defines personal data in Section1(1) as data which relate to a living individual who can be identified from those data or from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller.

This provides a wide definition of personal data, and it can be argued that all elements of our model are covered by this definition. This includes second-level data that are ascribed to an individual, by an organisation, as a result of analytics, for example, that based on the use of a specific item such as a model of iPhone. In the case of Durant v. Financial Services Authority [2003] EWCA Civ 1746, Auld LJ, the judgement limited personal data to that which affects a data subject's privacy, such as the subject's name, address, telephone coordinates, working interests and hobbies. In this interpretation, only the core of our model, the *digital mosaic*, is considered private data. However, following this judgement, the Information Commissioner's Office issued guidelines on the determination of personal data [76] in order to reconcile the Durant judgement with wider opinion. This lists eight questions, a positive reply to any of which may indicate that the information constitutes personal information. Once again, we see the defined boundaries of personal data expand out from the centre of our model to include the *digital persona* – data that can be used to inform or influence actions or decisions affecting an identifiable individual. The guidelines also include data that are linked to an individual. It therefore can be argued that second-level data at the edge of our model are also, under this definition, to be considered personal data. While this topic is more nuanced than we have shown here and deserves fuller analysis in a separate paper, we have demonstrated that the model can be used to illustrate the movement in the debate of what personal data are and, if accepted as a basis for legislation, could be used to define the boundaries of personal data.

7. Conclusion

The use of personal data continues to be a question of great interest in a wide range of fields, especially with respect to privacy, informed consent and right of access to data, driving a need for well-defined and consistent terms to describe those data. However, at present, the terminology around personal data is confusing, comprising multiple overlapping terms, with little agreement on the underlying concepts and their relationships.

In this article, we have presented a lexicological analysis of the terms used to describe personal data, based on an analysis of 247 papers (taken from an original sample of 64,584), and identified five distinct concepts (which we have labelled footprints, third-party footprints, mosaic, personas and extended self). These come together in a model of the *digitally extended self*. We have validated the model in two ways: by showing how 45 examples of usage from the literature map to the model (showing that each of the categories appears in the literature, although the terminology for them is inconsistent), and through case studies of an individual's real relationship with a financial institution, an international charity and a credit reference company, with data identified via subject access requests that illustrate how data held by these institutions fall into each category of the model.

The model of the digitally extended self that we have constructed is centred around the individual as a consequence of the overall context of personal data. The model shows that as data become more distant from the individual (moving from footprints to mosaics to personas), the questions of ownership, access and control of that data become less clear as it increasingly incorporates data from third parties (both individuals and organisations, in the form of their computer systems).

Our intention is to explore how the model can help with personal data transparency by extending our case study to the broader set of organisations and using these to analyse the quality of the data returned and the issues that organisations face in returning data at different layers of the model.

In addition, we anticipate that the categorisation model should prove to be particularly valuable to systems designers, as it establishes an overarching model for personal data. Computer scientists might also benefit from using it when examining the data constructs that may aid in the re-identification of individuals from theoretically anonymous data. Finally, legislators might value the definitions of categories of personal data that they provide to assist in debates regarding the boundaries of privacy law.

Our hope is that the model will enable discourse to continue on a common basis, facilitate a more focused debate and bring a better understanding of the relationships and structures inherent within personal data.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

Note

1. In the United Kingdom, under the 1998 Data Protection Act, an individual has the right, subject to certain exceptions, to get a copy of information that is held about them. A request for these data is known as a subject access request.

References

[1] Spiekermann S, Acquisti A, Böhme R et al. The challenges of personal data markets and privacy. *Electron Markets* 2015; 25: 161–167.
[2] Acquisti A, Brandimarte L and Loewenstein G. Privacy and human behavior in the age of information. *Science* 2015; 347: 509–514.
[3] Kehr F, Kowatsch T, Wentzel D et al. Blissfully ignorant: the effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Inform Syst J* 2015; 25: 607–635.
[4] Heeney C. Breaching the contract? Privacy and the UK census. *Inform Soc* 2012; 28: 316–328.
[5] L'Hoiry XD and Norris C. The honest data protection officer's guide to enable citizens to exercise their subject access rights: lessons from a ten-country European study. *Int Data Priv Law* 2015; 5: 190–204.
[6] Bakshi H. How can we measure the modern digital economy? *Significance*. Epub ahead of print 6 June 2016. DOI: 10.1111/j.1740-9713.2016.00909.x.
[7] Heinderyckx F. Reclaiming the high ground in the age of onlinement: ICA presidential address. *J Commun* 2014; 2014(64): 999–1014.
[8] Safran C, Bloomrosen M, Hammond WE et al. Toward a national framework for the secondary use of health data: an American medical informatics association white paper. *J Am Med Inform Assoc* 2007; 14: 1–9.
[9] Solove DJ. Introduction: privacy self-management and the consent dilemma. *Harvard Law Rev* 2013; 126: 1880–1903.
[10] Nissenbaum HF. *Privacy in context: technology, policy, and the integrity of social life*. Stanford, CA: Stanford Law Books, 2010.
[11] Pollach I. What's wrong with online privacy policies. *Commun ACM* 2007; 50: 103–108.
[12] Millard C and Hon WK. Defining 'personal data' in e-social science. *Inform Commun Soc* 2012; 15: 66–84.
[13] Polonetsky J, Tene O and Finch K. Shades of gray: seeing the full spectrum of practical data de-identification. *Santa Clara Law Rev* 2016, http://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=2827&context=lawreview
[14] Rowley J. The wisdom hierarchy: representations of the DIKW hierarchy. *J Inform Sci* 2007; 33: 163–180.
[15] Batra S. Big data analytics and its reflections on DIKW hierarchy. *Rev Manage* 2014; 4: 5.
[16] Palfrey J and Gasser U. *Born digital: understanding the first generation of digital natives*. New York: Basic Books, 2008.
[17] Craswell G and Poore M. *Writing for academic success*. London: SAGE, 2012.
[18] Bryman A. *Social research methods*. Oxford: Oxford University Press, 2012.
[19] Haddaway NR, Collins AM, Coughlin D et al. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PLoS ONE* 2015; 10: e0138237.
[20] Armstrong D, Gosling A, Weinman J et al. The place of inter-rater reliability in qualitative research: an empirical study. *Sociology* 1997; 31: 597–606.
[21] Glushko RJ, Annechino R, Hemerly J et al. Categorization: describing resource classes and type. In: Glushko RJ (ed.) *The discipline of organizing*. Cambridge, MA: MIT Press, 2013, pp. 235–272.
[22] Jones M. Super conducting super collider: evolution of facility layout requirement and CAD system development. In: *Proceedings of the unique underground structures symposium*, 1990, http://www.osti.gov/scitech/servlets/purl/6148417-PFSjrN/ (accessed 3 December 2014).
[23] Chang AYC, Parrales MEP, Jimenez J et al. Combining Google earth and GIS mapping technologies in a dengue surveillance system for developing countries. *Int J Health Geograph* 2009; 8: 49–59.
[24] Ludwig LF, Lauwers JC, Lantz KA et al. *Multimedia collaboration system with separate data network and a/v network controlled by information transmitting on the data network*. Patent US5617539 A, 1997.
[25] Goffman E. Role distance. In: Brissett D and Edgley C (eds) *Life as theater: a dramaturgical sourcebook*. New York: Aldine de Gruyter, 1990, pp. 101–111.
[26] Metzinger T. *The ego tunnel: the science of the mind and the myth of the self*. New York: Basic Books, 2010.
[27] Valk FV. Identity, power, and representation in virtual environments. *Merlot J Online Learn Teach* 2008; 4: 205–211.

[28] Byron T. *Safer children in a digital world: the report of the Byron review: be safe, be aware, have fun*. London: Department for Children, Schools and Families, 2008.
[29] Greysen SR, Kind T and Chretien KC. Online professionalism and the mirror of social media. *J Gen Int Med* 2010; 25: 1227–1229.
[30] Sellen A, Rogers Y, Harper R et al. Reflecting human values in the digital age. *Commun ACM* 2009; 52: 58–66.
[31] Lyon D. *Jesus in Disneyland: religion in postmodern times*. Cambridge: Polity Press, 2000.
[32] Turkle S. Constructions and reconstructions of self in virtual reality: playing in the MUDs. *Mind Culture Activity* 1994; 1: 158–167.
[33] Mcinnerney JM and Roberts TS. Online learning: social interaction and the creation of a sense of community. *Educ Techn Soc* 2004; 7: 73–81.
[34] Turkle S. Cyberspace and identity. *Contemp Sociol* 1999; 28: 643–648.
[35] Pearson E. All the World Wide Web's a stage: the performance of identity in online social networks. *First Monday* 2009; 14: 1–6.
[36] Bessière K, Seay AF and Kiesler S. The ideal elf: identity exploration in World of Warcraft. *Cyberpsychol Behav* 2007; 10: 530–535.
[37] Lee KM. Presence, explicated. *Commun Theory* 2006; 14: 27–50.
[38] Siibak A. Constructing the self through the photo selection – visual impression management on social networking websites. *Cyberpsychol* 2009; 3(1): article 1, http://www.cyberpsychology.eu/view.php?cisloclanku=2009061501&;article=1 (accessed 1 December 2014).
[39] Batchelor R, Bobrowicz A, Mackenzie R et al. Challenges of ethical and legal responsibilities when technologies' uses and users change: social networking sites, decision-making capacity and dementia. *Ethics Inform Tech* 2012; 14: 99–108.
[40] Wittes B. *Databuse: digital privacy and the mosaic*. Brookings Institute, 2011, https://www.technologylawdispatch.com/wp-content/uploads/sites/26/2011/05/GRE-Blog-May-17-2011-3.pdf (accessed 29 April 2017).
[41] Clark JE. The digital imperative: making the case for 21st century pedagogy. *Comput Compos* 2010; 27: 27–35.
[42] Siemens G and Long P. Penetrating the fog: analytics in learning and education. *Educ Rev* 2011; 46: 31–40.
[43] Chretien KC, Greysen SR, Chretien JP et al. Online posting of unprofessional content by medical students. *JAMA* 2009; 302: 1309–1315.
[44] Hankin C. *Foresight future identities 2013: final project report*. London: The Government Office for Science, 2013.
[45] Hengstler J. Managing your digital footprint: ostriches v. eagles. *Educ Digit World* 2011; 1: 89–139.
[46] Madden M, Fox S, Smith A et al. *Digital footprints: online identity management and search in the age of transparency*. Washington, DC: Pew Internet & American Life Project, 2007.
[47] O'Keeffe GS and Clarke-Pearson K. The impact of social media on children, adolescents, and families. *Pediatrics* 2011; 127: 800–804.
[48] Kapadia A, Henderson T, Fielding J et al. Virtual walls: protecting digital privacy in pervasive environments. In: *Proceedings of the 5th international conference on pervasive computing*, 2007, http://www.cs.dartmouth.edu/~dfk/papers/kapadia-walls.pdf
[49] Moore K and McElroy JC. The influence of personality on Facebook usage, wall postings, and regret. *Comput Human Behav* 2012; 26: 267–274.
[50] Briggs P. *Future identities: changing identities in the UK – the next 10 years. DR 4: will an increasing element of our identity be 'devolved' to machines?* London: The Government Office for Science, 2013.
[51] Gelman L. Privacy, free speech, and blurry-edged social networks. *Boston Coll Law Rev* 2009; 50: 1315–1344.
[52] Ess C. *Digital media ethics*. Cambridge: Polity Press, 2009.
[53] DeLillo D. *Mao II*. New York: Viking, 1991.
[54] Floridi L. Word of mouse. *Philos Mag* 2006; 33: 17.
[55] Westin AF and Ruebhausen OM. Privacy and freedom. New York: Atheneum, 1967.
[56] Garfinkel S. *Database nation: the death of privacy in the 21st century*. Sebastopol, CA: O'Reilly Media, Inc., 2000.
[57] Floridi L. The ontological interpretation of informational privacy. *Ethics Inform Tech* 2005; 7: 185–200.
[58] Smithson M. Toward a social theory of ignorance. *J Theory Social Behav* 1985; 15: 151–172.
[59] Solove DJ. Privacy and power: computer databases and metaphors for information privacy. *Stanford Law Rev* 2001; 53: 1393–1462.
[60] Ploeg IVD. Biometrics and the body as information. In: Lyon D (ed.) *Surveillance as social sorting: privacy, risk, and digital discrimination*. London: Routledge, 2003, pp. 57–73.
[61] Cherry S. Total recall (life recording software). *IEEE Spectrum* 2005; 42: 24–30.
[62] Gross R and Acquisti A. Information revelation and privacy in online social networks. In: *Proceedings of the ACM workshop on privacy in the electronic society*, 2005, pp. 71–80, https://www.heinz.cmu.edu/~acquisti/papers/privacy-facebook-gross-acquisti.pdf
[63] Solove DJ. *The digital person: technology and privacy in the information age*. New York: New York University Press, 2004.
[64] Dennis ES. Mosaic shield: Maynard, the fourth amendment, and privacy rights in the digital age. *Cardozo Law Rev* 2011; 33: 738–771.

[65] Hanna R, Rohm A and Crittenden VL. We're all connected: the power of the social media ecosystem. *Business Horizon* 2011; 54: 265–273.
[66] Schwartau W. *Information warfare: chaos on the electronic superhighway*. New York: Thunder's Mouth Press, 1994.
[67] Clarke R. Web 2.0 as syndication. *J Theor Appl Electron Commerce Res* 2008; 3: 30–43.
[68] Solove DJ. A taxonomy of privacy. *Univ Pennsylvania Law Rev* 2006; 154: 477–564.
[69] Andrews L. *I know who you are and I saw what you did: social networks and the death of privacy*. New York: Free Press, 2013.
[70] Ardagna CA, Camenisch J, Kohlweiss M et al. Exploiting cryptography for privacy-enhanced access control: a result of the prime project. *J Comput Security* 2010; 18: 123–160.
[71] Blanchette JF and Johnson DG. Data retention and the panoptic society: the social benefits of forgetfulness. *Inform Soc* 2002; 18: 33–45.
[72] Hildebrandt M and Gutwirth S. *Profiling the European citizen: cross-disciplinary perspectives*. Berlin: Springer Science & Business Media, 2008.
[73] Clarke R. Computer matching and digital identity. In: *Proceedings of the ACM computers, freedom & privacy conference*, Burlingame, CA, 9–12 March 1993. New York: ACM.
[74] Ludington S. Reining in the data traders: a tort for the misuse of personal information. *Maryland Law Rev* 2006; 66: 140–193.
[75] Sanchez A. Facebook feeding frenzy: resistance-through-distance and resistance-through-persistence in the societied network. *Surveill Soc* 2009; 6: 275–293.
[76] Information Commissioner's Office. *Determining what is personal data*. 2012, https://ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf (accessed 15 March 2017).