Article

The Digitally Extended Self: A Lexicological Analysis of Personal Data

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

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

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

Richard Giordano
Faculty of Health Science, University of Southampton

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 paper, 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, that extend out from the individual and form the Digitally Extended Self.

1. Introduction

The continuing concerns about the use of personal data, especially with respect to privacy, informed consent, and the right of access to data, drive a need for well-defined and consistent terms to describe that data. This paper 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], have 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 be related to each other? A variety of terms present themselves in the literature, for example digital 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

Corresponding author:
Brian Parkinson, University of Southampton, 71 Bainton Road, Oxford, OX2 7AG,
Email: BLP1M11@soton.ac.uk

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 ICT terminology. Others have tried to address the problem in other domains, for example Safran [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. Additionally, 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 analyze 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 paper is arranged as follows: Section 2 describes 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 paper 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 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 inconsistences 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 is 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 it can be transformed into information, knowledge or wisdom, each having the potential to be more threatening to an individual's privacy. Whilst 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 & 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, whilst 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, e.g. 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 1,000 search results [19] was 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*), these then became our seed search terms for Google Scholar. The search engine, at the time of this work, returned a maximum of 1,000 items for each search and so for high usage terms we searched by calendar year thus maximizing 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 below in Table 1 together with the total count of results, and the number of papers selected under each term for the purposive sample).

Table 1. Summary of Google Scholar Search Results, Aug 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	25578	12
Digital Mosaic	834	11	Shadow Identity	171	2
Digital Person	967	39	Virtual Self	5120	43

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 usage, each term was taken in turn, and the sample documents, containing that term, examined. Meanings were observed and common themes extracted. The terminology descriptive of personal data was then examined and through a series of iterations, involving the second author, a standard categorisation 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] notes, 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 debate between the authors with the objective of producing a consistent set of terms that can be used when discussing personal data. Whilst others may have decided on an alternate nomenclature we have endeavoured to create a categorization 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 sixteen noun phrases, and within a discourse, the meanings tend to cover overlapping sets of things, as we will present below. However, when analysing 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, whilst Metzinger [26] uses the term to cover phantom limbs, dream states and out of body experiences, and Valk [27] considers 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 is 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] p. 33, amongst others, state that

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

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] p. 166 sees as

> extensions of ourselves we have embodied in program.

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, whilst Bessière, Seay, and Kiesler [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 else 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, e.g. 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 below 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 artifacts 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 whilst 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 artifact 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:

- *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 provides information about an individual's attributes (e.g. demographic data which is associated with a person's post/ZIP code).

There are several reasons for placing these categories together in a coherent model. Firstly, 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. Secondly, 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 we affect how they are viewed. 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 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 the analysis. The whole, is then defined as the *Digitally Extended Self.*
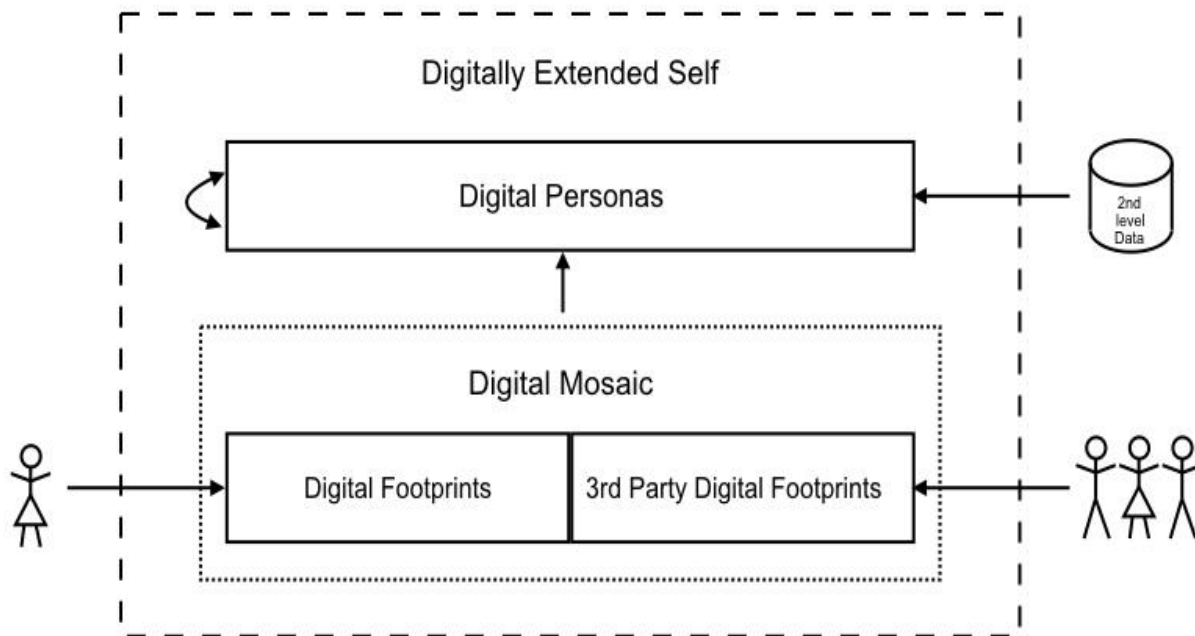

Figure 1. Our model of the Digitally Extended Self – showing the five categories of personal data.

6. Validation of the Model

The model serves two purposes. It provides a clear nomenclature which facilitates a cross-disciplinary use of terms, 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 disorganized, terminology; and second, against actual data.

6.1. Validation against terminology

As a first validation step we show below (Tables 2 through 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 (for example, Byron [28] discusses digital footprints in the same way as our model, but Chretien et al. [42] 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 the tables below.

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.

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' p2	Wittes [40]
	2.2 digital footprint	2.2.1 a digital artifact left behind by some activity 'as they "tread" through the World Wide Web, they leave behind a "footprint"' p1227	Batchelor et al. [39] Siemens & Long [43] Greysen et al. [29]
		2.2.2 'personal information available online' p58	Byron [28]
		2.2.3 postings on social media (by medical students)	Chretien et al. [42]
		2.2.4 patterns of internet usage / artifacts	Hankin et al. [44]
		2.2.5 traces of online presence	Hengstler [45] Madden et al. [46] O'Keeffe & 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 & Gasser [16]
		2.2.8 a group of digital footprints on one site i.e. Facebook	Moore & 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' p10	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. [42] Ess [52]
	3.3 digital mosaic	3.2.1 'He was a digital mosaic … storing his data in starfish satellites' p112	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' p2	Wittes [40]

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 artifacts	Westin [55] Garfinkel [56] Floridi [57]
		4.1.2 'records and data about the self' p167	Smithson [58]
	4.2 digital biography	4.2.1 'an electronic collage' p1394, 'a life captured in records' p1394, 'bits and pieces of stored information about one's life' p70	Solove [59] Ploeg [60]
	4.3 digital doppelganger	4.3.1 a collection of digital artifacts which provide a picture of a life	Cherry [61]
	4.4 digital dossier	4.4.1 collections of footprints e.g. from Facebook	Gross & Acquisti [62] Garfinkel [56] Solove [63]
	4.5 digital footprint	4.5.1 digital artifacts 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 publically available	Palfrey & 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 artifacts 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' p1	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 publically available	Palfrey & 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' p70	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 & Johnson [71]
		5.3.3 personas derived from profiling and data mining	Hildebrandt & 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' p1	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' p142	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' p39	Palfrey & Gasser [16]
	6.3 digital persona	6.3.1 analysis of transactional data combined with other records e.g. demographics	Blanchette & Johnson [71]
	6.4 virtual self	6.4.1 analyses computed by marketing companies and government departments augmented by further transactions	Lyon [31]

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 a 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 auto ethnography 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 central and local government, public and private companies, and 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 the first author, 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 to 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, 3 and 4 show 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 artifacts 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 the *digital extended self*.
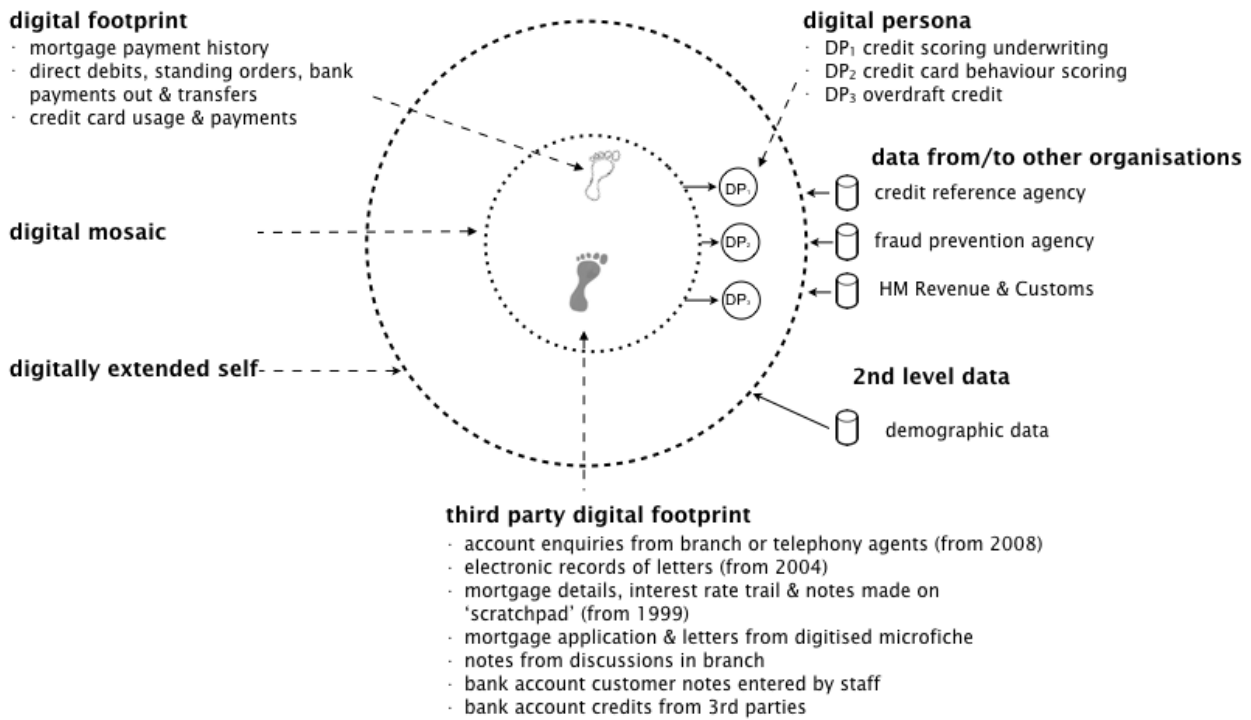
digital footprint
- mortgage payment history
- direct debits, standing orders, bank payments out & transfers
- credit card usage & payments

digital mosaic

digitally extended self

digital persona
- DP$_1$ credit scoring underwriting
- DP$_2$ credit card behaviour scoring
- DP$_3$ overdraft credit

data from/to other organisations
credit reference agency
fraud prevention agency
HM Revenue & Customs

2nd level data
demographic data

third party digital footprint
- account enquiries from branch or telephony agents (from 2008)
- electronic records of letters (from 2004)
- mortgage details, interest rate trail & notes made on 'scratchpad' (from 1999)
- mortgage application & letters from digitised microfiche
- notes from discussions in branch
- bank account customer notes entered by staff
- bank account credits from 3rd parties

Figure 2. The centric diagram showing data (as instances of the model) from case study 1, a UK based bank.
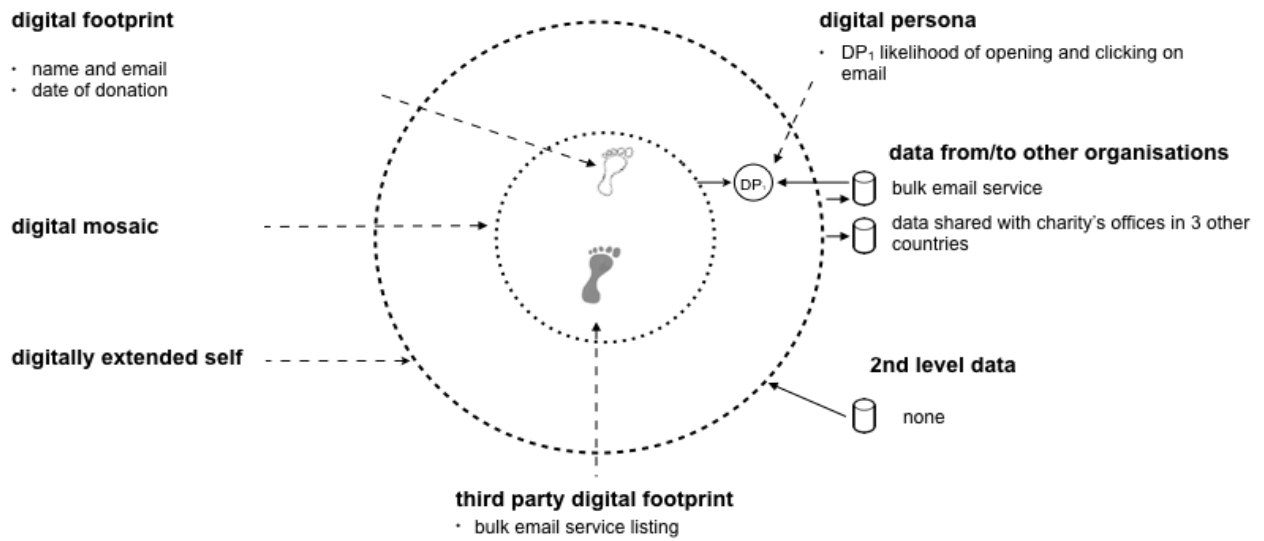
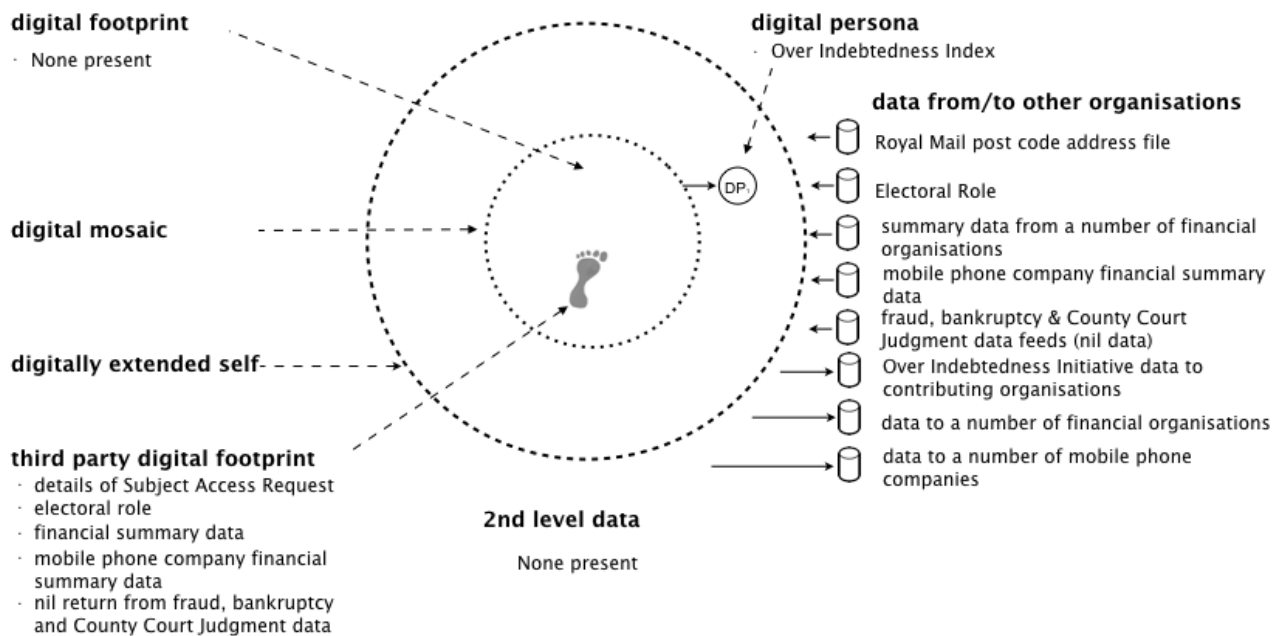**Figure 3.** Case study 2, an international charity.

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

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 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 describes 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 even though a minimal level of data was held by this organization, (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 is sent to the charity's offices in three other countries, two of which are approved by the 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 organization. This private limited company had no direct contact with the data subject but collected third party digital footprints in the form of summary data from financial and mobile phone companies, which it combined with post code, electoral role, fraud, bankruptcy, and court judgment data. This information was used to calculate an over indebted assessment which, together with other data, were 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 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 becomes less direct and more speculative as the data becomes 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 UK, there has been an ongoing 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 is ascribed to an individual, by an organization, 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 co-ordinates, working interests or 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 judgment 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 is linked to an individual. It therefore can be argued that second level data at the edge of our model is also, under this definition, to be considered personal data. Whilst 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 is, 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 the right of access to data, driving a need for well-defined and consistent terms to describe that data. However, at present the terminology around personal data is confusing, comprised of multiple overlapping terms, with little agreement on the underlying concepts and their relationships.

In this paper, 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, even though 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 falls 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 becomes 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 it provides 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.

Notes

1. In the UK, 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.

Funding

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, 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. *Information Systems Journal* 2015; 25: 607-635.
[4] Heeney C. Breaching the contract? Privacy and the UK census. *The Information Society* 2012; 28: 316-328.
[5] L'hoiry XD, 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. *International Data Privacy Law* 2015; 5: 190-204.
[6] Bakshi H. How can we measure the modern digital economy? *Significance* 2016; 6-7.
[7] Heinderyckx F. Reclaiming the high ground in the age of onlinement: ICA presidential address, 2014. *Journal of Communication* 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. *Journal of the American Medical Informatics Association* 2007; 14: 1-9.
[9] Solove DJ. Introduction: Privacy self-management and the consent dilemma. *Harvard Law Review* 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. *Communications of the ACM* 2007; 50: 103-108.
[12] Millard C, Hon WK. Defining 'personal data' in e-social science. *Information, Communication & Society* 2012; 15: 66-84.
[13] Polonetsky J, Tene O, Finch K. Shades of gray: Seeing the full spectrum of practical data de-identification. *Santa Clara Law Review, Forthcoming* 2016;
[14] Rowley J. The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science* 2007; 33: 163-180.
[15] Batra S. Big data analytics and its reflections on DIKW hierarchy. *Review of Management* 2014; 4: 5.
[16] Palfrey J, Gasser U. Born digital: Understanding the first generation of digital natives. New York, NY: Basic Books, 2008.
[17] Craswell G, 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:
[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 organising. 2013, p. 235-272.
[22] Jones M. Super conducting super collider: Evolution of facility layout requirement and cad system development. *Unique Underground Structures Symposium* 1990; http://www.osti.gov/scitech/servlets/purl/6148417-PFSjrN/ (accessed December 3 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. *International Journal of Health Geographics* 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. *Vicor, Inc, Palo Alto, Calif* 1997; U.S.A.:
[25] Goffman E. Role distance. In: Brissett D, Edgley C, (eds). Life as theater: A dramaturgical sourcebook. New York: Aldine de Gruyter, 1990, p. 101-111.
[26] Metzinger T. The ego tunnel: The science of the mind and the myth of the self. New York, NY: Basic Books, 2010.

[27]   Valk FV. Identity, power, and representation in virtual environments. *Merlot Journal of Online Learning and Teaching* 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. Department for Children, Schools and Families, 2008.

[29]   Greysen SR, Kind T, Chretien KC. Online professionalism and the mirror of social media. *Journal of General Internal Medicine* 2010; 25: 1227-1229.

[30]   Sellen A, Rogers Y, Harper R et al. Reflecting human values in the digital age. *Communications of the 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, and Activity* 1994; 1: 158-167.

[33]   Mcinnerney JM, Roberts TS. Online learning: Social interaction and the creation of a sense of community. *Educational Technology & Society* 2004; 7: 73-81.

[34]   Turkle S. Cyberspace and identity. *Contemporary Sociology* 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 [Online]* 2009; 14:

[36]   Bessière K, Seay AF, Kiesler S. The ideal elf: Identity exploration in World of Warcraft. *Cyberpsychology & Behavior* 2007; 10: 530-535.

[37]   Lee KM. Presence, explicated. *Communication Theory* 2006; 14: 27-50.

[38]   Siibak A. Constructing the self through the photo selection - visual impression management on social networking websites. *Cyberpsychology : Journal of Psychosocial Research on Cyberspace* 2009; 3(1): Article 1.http://www.cyberpsychology.eu/view.php?cisloclanku=2009061501&article=1 (accessed December 1 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 and Information Technology* 2012; 14: 99-108.

[40]   Wittes B. Databuse: Digital privacy and the mosaic. *Brookings Institute* 2011; http://www.globalregulatoryenforcementlawblog.com/uploads/file/GRE Blog - May 17, 2011 - 3.pdf (accessed July 2 2013)

[41]   Clark JE. The digital imperative: Making the case for 21st century pedagogy. *Computers and Composition* 2010; 27: 27-35.

[42]   Chretien KC, Greysen SR, Chretien JP et al. Online posting of unprofessional content by medical students. *Jama: The Journal of the American Medical Association* 2009; 302: 1309-1315.

[43]   Siemens G, Long P. Penetrating the fog: Analytics in learning and education. *Educause Review* 2011; 46: 31-40.

[44]   Hankin, C et al. Foresight future identities (2013) final project report. London: The Government Office for Science, 2013.

[45]   Hengstler J. Managing your digital footprint: Ostriches v. Eagles. *Education For A Digital 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, 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. *5th International Conference on Pervasive Computing* 2007;

[49]   Moore K, Mcelroy JC. The influence of personality on Facebook usage, wall postings, and regret. *Computers in Human Behavior* 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 College Law Review* 2009; 50: 1315-1344.

[52]   Ess C. Digital media ethics. Cambridge: Polity Press, 2009.

[53]   DeLillo D. Mao II. New York, NY: Viking, 1991.

[54]. Floridi L. Word of mouse. *The Philosophers Magazine* 2006; 33: 17.

[55]   Westin AF, Ruebhausen OM. Privacy and freedom. New York, NY: 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 and Information Technology* 2005; 7: 185-200.

[58]   Smithson M. Toward a social theory of ignorance. *Journal for the Theory of Social Behaviour* 1985; 15: 151-172.

[59]   Solove DJ. Privacy and power: Computer databases and metaphors for information privacy. *Stanford Law Review* 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, p. 57-73.

[61]   Cherry S. Total recall (life recording software). *Spectrum, IEEE* 2005; 42: 24-30.

[62]   Gross R, Acquisti A. Information revelation and privacy in online social networks. *ACM Workshop on Privacy in the Electronic Society* 2005; 71-80.

[63]   Solove DJ. The digital person: Technology and privacy in the information age. New York, NY: New York University Press, 2004.

[64] Dennis ES. Mosaic shield: Maynard, the fourth amendment, and privacy rights in the digital age. *Cardozo Law Revue* 2011; 33: 738 - 771.

[65] Hanna R, Rohm A, Crittenden VL. We're all connected: The power of the social media ecosystem. *Business Horizons* 2011; 54: 265-273.

[66] Schwartau W. Information warfare: Chaos on the electronic superhighway. New York, NY: Thunder's Mouth Press, 1994.

[67] Clarke R. Web 2.0 as syndication. *Journal of Theoretical and Applied Electronic Commerce Research* 2008; 3: 30-43.

[68] Solove DJ. A taxonomy of privacy. *University of Pennsylvania Law Review* 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. NY, NY: 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. *Journal of Computer Security* 2010; 18: 123-160.

[71] Blanchette JF, Johnson DG. Data retention and the panoptic society: The social benefits of forgetfulness. *The Information Society* 2002; 18: 33-45.

[72] Hildebrandt M, Gutwirth S. Profiling the European citizen: Cross-disciplinary perspectives. Berlin: Springer Science, 2008.

[73] Clarke R. Computer matching and digital identity. *Proceedings of the ACM Computers, Freedom & Privacy Conference* 1993;

[74] Ludington S. Reining in the data traders: A tort for the misuse of personal information. *Maryland Law Review* 2006; 66: 140-193.

[75] Sanchez A. Facebook feeding frenzy: Resistance-through-distance and resistance-through-persistence in the societied network. *Surveillance & Society* 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)