This is the accepted manuscript of the article, which has been published in Archival Science. 2017, 17(3), 285–303. http://dx.doi.org/10.1007/s10502-017-9277-0

Privacy as an archival problem and a solution

Abstract

People feel that their privacy is violated when information about them is passed inappropriately from one context or social sphere to another. This makes records and archives management a focal point of privacy issues, because its goal is to transfer information from one context, place, and point in time to other contexts, places, and points in time. Society has a number of mechanisms ("strategies") for protecting the privacy of people. The article examines five of them (purpose limitation, privacy self-management and right to be forgotten, destruction, anonymization, and information safe haven approach) and the limits they set to the contextual transfer of information. If the strategies are implemented in society without regard to archival needs, archives have difficulties in fulfilling their functions in society. Therefore, records professionals should make their point of view known when privacy issues are discussed. Records professionals also should be aware of the mutability of the category of personally identifiable information and the changing nature of privacy issues in the digital environment.

Introduction

It is impossible to imagine an area in records and archives management that has not been transformed by digitalization: by the adoption of digital or computer technology in the society and the way many domains of social life are restructured around digital communication and media infrastructures (Brennen and Scott 2014). Digitalization has changed the way information is created, managed, preserved, disseminated, and used. This has changed our information environment and the nature of privacy problems. Digitalization has made privacy a burning issue.

This paper examines what has happened to privacy in the digital environment and how it affects archives. Traditional archival processes were adapted to working with paper documents, but today—and increasingly in future—challenges in protecting privacy are different from what they were some decades ago. Although the literature on privacy in the digital environment is vast, most writers ignore the archival point of view. Therefore, it is important to look at what has been written about privacy in the changing information environment and put it in relationship with the archival mission—which is to preserve information that has been found worthy of preservation by making impartial reasoned judgments about its value for future generations. My argument is that the way the society protects the privacy of people is crucial for archives. It influences what information is generated, how it is stored and how it can be accessed. At the same time, archival techniques—and recordkeeping techniques in general—are critical for the protection of personal privacy now and in the future.

In this article, the first section discusses how digitalization has changed the questions of privacy. The second section argues that archives are in the focal point when it comes to privacy because they carry information between contexts. The third section examines societal strategies which are common in Western societies for protecting privacy from the archival perspective: binding to purpose, privacy self-management and right to be forgotten, destruction, anonymization, and information safe haven approach. Finally, the paper suggests areas that records professionals should follow closely and in which they should be active.

Published in Archival Science (2017). The final publication is available at Springer via http://dx.doi.org/10.1007/s10502-017-9277-0

Changing form of privacy issues

Article 12 of The Universal Declaration of Human Rights states

> "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks." (United Nations 1948)

Underpinning the 12th article one can see some of the ideas that have formed the backbone of traditional thinking about privacy. First is the idea that there is a private space distinct from the public space. One needs protection in the private space: it is the space that must be protected from arbitrary interference and attacks (Nissenbaum 1998). Secondly, a general assumption has been that the source of attacks to privacy is the state (Duchein 1983, p 19; MacNeil 1992, pp 35–36). Thirdly, there are categories of information that belong to the sphere of privacy (like civil status and affiliation, health, wealth and income, political and religious opinions, family honor, and so on) (Duchein 1983, p 20). Although the list of the categories that are considered sensitive varies from one country to another it is assumed that we can identify sensitive and non-sensitive information by looking at it.

Today all these assumptions are inadequate if not entirely wrong. Privacy threats posed by computerized information processing have been evident for a long time (see Hedstrom 1981), but the development of the networked society has exacerbated them. Firstly, while governments and security organizations collect more information about citizens than ever before they are not the sole threat to privacy. In addition to the state, private enterprises gather vast amounts of information. This information can be on-sold to a new party which may combine it with other information from public and non-public sources.

Even more important for archives is what has happened to the other traditional assumptions about privacy: the idea that we can separate both private and public space and private and non-private information. The borderlines—if they exist—are no more self-evident. The capacity of society to record, recall and process information has exponentially grown with digitalization. Although the change is in many ways more quantitative than qualitative it has destroyed the distinction between public and private space and private and public information.

Traditionally, it was thought that a person needs protection in the private space, but not in the public space. What happened in the public space could be observed by everyone and distributing information about it could not be circumscribed without limiting others' freedom of expression. However, in practice the capability for observation and storing of information was limited, and a person in a public space—in a busy street, for instance—could be observed by all and still be unnoticed by everyone. Today, every action that takes place on the internet can be tracked, saved, and analyzed indefinitely. What happens is never missed by those who want observe it. Information can be kept in memory and shared with new audiences almost without a limit (Nissenbaum 1998). "Since the beginning of time, for us humans, forgetting has been the norm and remembering the exception", notes Viktor Mayer-Schönberger. Now, "…because of digital technology, society's ability to forget has become suspended,

replaced by perfect memory." (Mayer-Schönberger 2011, pp 2, 4.) What takes place in public can be kept in a memory indefinitely.

Private space has experienced a similar transformation. What happened in the private space was previously confined between four walls and subject to mis-hearing and human forgetfulness. Today smart phones and gadgets like Google Glass make it possible to record what happens in private and further distribute this information immediately. In addition, "secondhand data leakage" is growing in prevalence, meaning that one person's action impacts another's private data (for example, when a friend declares a co-location with us, or a blood relative unveils her genome) (Hubaux and Juels 2016). Big data involves not only individuals' digital footprints (data they themselves leave behind), but, perhaps more importantly, also individuals' data shadows (information about them generated by others) (Koops 2011). This means that the walls protecting private space are crumbling. All spaces are potentially public.

Furthermore, exponential growth in the society's capacity to record, recall and process information is destroying any distinction between private and non-private information. One may argue that it has always been a borderline drawn in water. Privacy can be understood as a self-evident quality inherent in documents only if it assumed that archives are mere repositories of inert documents whose meaning is fixed and fully-present. This proposition has been challenged and questioned for a long time now. Restrictions protecting private documents are themselves conditions that are built in part of documents' meaning, interpretation and value (Dever 2012). In addition, while privacy is seemingly a universal concept, what is private and under what circumstances differs widely between groups of people and over time. For instance, in the United States even the archivist appraising records of the FBI was not allowed to see income tax return information in the files, but in Sweden, by contrast, personal income tax information is open to public immediately (Duchein 1983, pp 19–22; Huskamp Peterson 2007). Thus, something becomes private when it is interpreted as private.

However, digitalization has made the distinction between private and non-private blurred in a new way. Everything can be tracked in the digital environment. When all the innocent and nothing-to-hide information about what we do in the public space—e.g. go shopping, make friendships, support a good cause, have discussions—is put together little room is left for the privacy even though any single item of information by itself would not form a privacy threat. This has two consequences. Firstly, protecting private space is not enough. People need protection also in the public space and protection from the gathering of seemingly innocuous pieces of information: they need "privacy in public", as Helen Nissenbaum (1998) has argued. Secondly, identifying information that threatens privacy is more difficult, because it is the accumulation of information that is decisive, rather than the individual facts.

Privacy and records and archives management

Ivan Szekely (2014) has argued that archives are now moving from the "public paradigm"—in which the archivist was the key professional giving access to the state, historians, and learned or concerned public in reading rooms and via public services—to the "global paradigm" where IT professionals and information brokers deliver global

services to all internet users. Records professionals should ask how the change of privacy issues in the digital environment affects their work. A disconcerting answer would be that it does not have any major implications. In that case, records professionals may have been superseded by other information management professionals—as Szekely seems to suggest—from the core of digital information management.

In my view records and archives management is even more closely intertwined with privacy issues than it may first seem. The connection is in part hidden because privacy is a theoretically challenging concept, although it seems straightforward at first glance. Privacy is like an elephant: easy to recognize, difficult to describe (Young 1978; MacNeil 1992, p 9). We do not have a good theory about what privacy is and why it matters so much to us (Schonsheck 1997).

Nevertheless, it seems that we cannot speak about privacy without also talking about contexts of information usage. It was already noted that any violation of privacy is in part a matter of interpretation and dependent on the context where the interpretation is made. Helen Nissenbaum (1998) emphasizes contextuality of privacy from another perspective. She states that a key to privacy is considering "contextual integrity". By this she means that we all have a sense of what is appropriate usage of our information. It feels acceptable that the doctor who is treating us has access to our medical records or that the merchant by whom we make our daily shopping knows what we have bought, but if this information is spread beyond this intuitively appropriate sphere we feel that our privacy has been violated. In short, we are willing to share information with others in a particular context and for a particular purpose—for which we have weighted pros and cons (Mayer-Schönberger 2011, p 101). Jonathan Schonscheck (1997) notes among others (see Vistilä and Ruokonen 2016) that we are members of a number of social spheres: there are close friends, spouse, nuclear family, extended family, and specialized spheres like "me and my banker", to name only few. It is essential that one can control what information is available to any of these spheres. (Schonsheck 1997.)

More crucial than the number of spheres is here the very idea that the privacy is being violated when information crosses borders of "contexts" or "social spheres". Records and archives management is at the focal point of privacy issues, because it exists precisely to transfer information in usable and understandable form from one context and point in time to another context and time. This idea repeats itself in different forms in records and archives management literature. For instance, in the life cycle model an organization first uses records to support its work. Thereafter they are transferred to an archival institution to be used by new user groups for purposes for which the records were not initially created. While the records continuum model, on the other hand, does not make a distinction between the active or semi-active phase of records (records management) and the historical phase (archives), it also states that records (and archives) serve several users and purposes in different contexts. The records continuum model shows how records and archives management consists of processes that make information available to ever larger user groups starting from the immediate neighborhood of information creation inside the organization and expanding from that to the whole organization and finally to the society at large. (For life cycle and records continuum models, see e.g. An 2003.) A third example: Recordkeeping

Metadata Working Meeting of the Dutch Archiefschool and Netherlands Institute for Archival Education and Research noted in year 2000 that recordkeeping metadata functions supports the transfer of records across domains and over time. Recordkeeping metadata was consequently defined as "structured or semi-structured information which enables the creation, management, and use of records through time and within and across domains in which they are created." (Hedstrom 2000, 2001.)

If records and archives management consists among other things of a set of processes which transfer and offer access to information across contextual, spatial, and temporal boundaries and the very crossing of boundaries itself creates privacy violations then issues of privacy are bound to rise. As Elena Danielson (2010) notes, “the violation of privacy is an intrinsic and unavoidable part of archival work, because it involves the secondary use of documents, which were originally created for another, so called primary, purpose.” However, the crossing of boundaries does not happen only when the information comes to an archive to be used for secondary purposes. It also happens when information is shared inside the organization, with interest groups, or delivered to supervisory bodies. While the transition of information into the sphere of secondary use represents a profound contextual change, as Danielson argues, smaller contextual transfers happen all the time. Also this shapes the records. If information does not remain “private” (in the sense that it would be known only by its creator or a small network of trusted parties) it shapes the record’s content and people’s behavior from the beginning. Something is perhaps left unrecorded or what is said is modified to fit the expected audience. All records presume an audience (Maanen and Pentland 1994, p 54). Records are often produced to document the performance of a given organizational task. In those cases, organizational members devote at least a part of their labor to the creation of the desired impression as expressed by means of documentation. If it is not known who is going to use the information and for what purpose, it is safest to assume the worst. In the digital world we are living in Bentham’s panopticon without spatial and temporal limits: we do not know if our current actions are being (or if they will be) watched, but it is better to be cautious and behave accordingly (Mayer-Schönberger 2011, pp 109-11).

For archives this creates a danger of a “chilling effect” or “empty archives” -syndrome: to avoid responsibility or blame information is sanitized or meetings conducted in ways (e.g. in telephone) that do not create a record (about Hillary Clinton’s refusal to keep diary, see Danielson 2010; Pasquier and Villeneuve 2007; Shepherd 2015; about the lack of empirical evidence for chilling effect, see Worthy 2010).

Societal strategies for protecting privacy

In the literature, one can identify several “societal strategies”—possible courses of action—for protecting the privacy of individuals. Society can 1) follow the purpose limitation principle 2) rely on information self-management and acknowledge the right to be forgotten, 3) destroy information, 4) anonymize information, or 5) create for personal information a safe haven where the information is kept protected until it can be opened for public access. The five strategies have been selected here for closer

examination because they have corollaries for records and archives management by setting limits to the contextual transfer of information.

The five strategies have a common feature in that you can find them in slightly different forms everywhere in legislation, standards, and textbooks of best practice. Strategy implementations may vary. The focus here is on the strategies themselves and their impact on the contextual transfer, not on the way a strategy can be implemented. While a strategy implementation, like building a digital privacy rights infrastructure, is also socially important, it does not change the underlying strategy itself. For instance, being transparent about what data is collected, why, and for what purpose ("fair processing") does not itself change the strategy and the limits it sets to the contextual transfer of information. Another example is the "MyData" concept. MyData refers to a shared technical infrastructure that empowers individuals as managers of data about them. Thus, it is a manifestation of the second strategy in which the person is given the power over personal information about himself.

Many of the societal strategies were first introduced in the code for Fair Information Practices (FIP). The code was created in year 1973 by the advisory committee of the US Department of Health, Education, and Welfare. The immediate impetus for the FIP was the increasing use of computers for the transaction of ordinary government and business transactions. In 1980 both the Organization for Economic Co-Operation and Development (OECD) and the Council of Europe published guidelines that relied on the FIP core principles. Although today few writers outside the circle of records professionals explicitly mention records when they write about the privacy issues, the original FIP code stated that its fundamental principles "require adherence of record-keeping organizations" (Gellman 2016.) Next the five strategies are discussed one by one.

Purpose limitation principle

The first societal strategy for protecting privacy is to follow the purpose limitation principle. In essence, the purpose limitation principle provides that any processing of personal information must be compatible with the purpose or purposes specified and expressed at the time of the collection of the information (Rauhofer 2014). Four international privacy instruments (the OECD guidelines, Council of Europe Convention, EU Data Protection Directive, and the Asian-Pacific Economic Cooperation (APEC) Privacy Framework) share today this idea (Greenleaf 2012). It has its origin in the basic FIP principles of 1973 (Gellman 2014). The effectiveness of the purpose limitation principle depends on the society's willingness to abide by present information privacy principles in the future (Mayer-Schönberger 2011, p 138).

Privacy self-management and right to be forgotten

Another alternative is to give the power over personal information to the person concerned. In this strategy, the person decides either about the usage of information (privacy self-management) or about the destruction of information (right to be forgotten). The ideas of privacy self-management and right to be forgotten are conceptually distinct and they do not have the same historical background. However, they are grouped here together because in both the subject decides what happens to the

information. The right to be forgotten is here seen as the ultimate form of privacy self-management: deleting data by the request of the individual ensures that no-one will ever have access to it.

The idea of privacy self-management is that people have control over their personal data and they evaluate themselves the costs and benefits of collection, use or disclosure of their information. Privacy self-management does not take a stance on whether certain forms of collecting, using, or disclosing personal information are good or bad. Nearly all forms can be legitimized by consent (Solove 2013).

Privacy self-management is problematic for many reasons. Firstly, there are severe cognitive problems that undermine privacy self-management:

> "(1) People do not read privacy policies; (2) if people read them, they do not understand them; (3) if people read and understand them, they often lack enough background knowledge to make an informed choice; and (4) if people read them, understand them, and can make an informed choice, their choice might be skewed by various decision making difficulties." (Solove 2013.)

Besides these cognitive problems that prevent people from making rational choices, there are structural problems as well: people might be able to manage privacy with a few entities, but the average American visits nearly a hundred websites per month and does business online and offline with countless companies. There are too many entities collecting and using personal data to make it feasible for people to manage their privacy separately with each unity. Moreover, information is collected over time which makes it virtually impossible for people to weigh the costs and benefits of giving up their privacy (Solove 2013.) Data aggregates slowly and bits of innocuous information can say a lot in combination. This makes it practically impossible to manage the data and make meaningful judgments about the costs and benefits of revealing it. Individual privacy breaches are often small and dispersed, but cumulative in nature. (Solove 2013, pp 1889–1891.) Gordon Hull (2015) says that effectuating privacy preferences is difficult also because the choice is not actually "free": social cost of being outside e.g. Facebook can be too high to bear. He also criticizes privacy self-management by arguing that it "isn't about protecting people's privacy; it's about inculcating the idea that privacy is an individual, commodified good that can be traded for other market goods" (Hull 2015).

The right to be forgotten means that a person can have digital data deleted so that third parties can no longer trace him (Weber 2011). The data subject has the right to obtain from the controller of data the erasure of personal data and the abstention from further dissemination of such data (Szekely 2014). This has only fairly recently been recognized as a fundamental right. There is no consensus on what exactly a right to be forgotten means, and its status is unclear: is it a right, interest, ethical or social value, or policy aim; is it in need of reinforcement or to be created from scratch (Koops 2011). In continental Europe the right to be forgotten is understood as being contained in the right of the personality, encompassing several elements such as dignity, honor, and the right to private life. (Weber 2011.) The right to be forgotten is "a right to have the government stop you from speaking about me". In the United States courts have usually favored the right to expression instead of right to be forgotten, unless sensitive information is disclosed after interventions into the private sphere have been done in

frivolous and socially irredeemable forays (Weber 2011). The right to be forgotten has raised the concern of subtle censorship and loss of information that is needed in future (Ausloos 2012).

Destruction

The simple existence of information about a person can be a threat to privacy, because it creates the risk that the person could be harmed later (Solove and Scott 2006, pp 487–488). Therefore, one solution for privacy problems is to destroy the information altogether. Some writers have argued for "institutionalized government amnesia": some data must be destroyed after a certain amount of time unless there are good reasons for retaining it. "Information ecology rules" explicitly regulating how long information can be retained are nothing new. (Mayer-Schönberger 2011, pp 158–9.) The idea of destroying information is often an extension to the purpose limitation principle: once the information is no longer needed for the original purpose, it must be destroyed. Also the right to be forgotten includes the idea of protecting privacy by deleting information permanently. However, information can be destroyed as a paternalistic measure without the request of the individual and regardless of the purpose of collection.

Anonymization

Privacy regulations are commonly triggered by the presence of "personally identifiable information". Privacy laws typically regulate only when personally identifiable information is involved. (Schwartz and Solove 2011, pp 1814, 1816, 1890.) Therefore, anonymization of data has become the favored technique in legislation, standards, and guidelines for best practice.

Once the data is anonymized, it contains no more personally identifiable information and can be shared with others. Anonymization seems to offer the best-of-both-worlds compromise: analysts will still find the data useful, but unscrupulous marketers and malevolent identity thieves will find it impossible to identify the people tracked. Nearly every information privacy law or regulation grants a get-out-of-jail-free card to those who anonymize their data. (Ohm 2010, pp 1703–4).

For decades technologists have believed that it is possible to robustly protect people's privacy by making small changes to data (Ohm 2010, pp 1703, 1706–7, 1714.) Anonymization of data is commonly done by omitting identifying database values—like names or social security identification numbers—entirely (data masking), by replacing the values with new artificial identifiers (pseudoanonymization), or by changing the values to a more generic form (by omitting the end of the zip code, for instance). There are several anonymization techniques that can be used either to produce aggregated information or anonymized data on an individual-level basis (for details, see e.g. ICO 2012).

Unfortunately, anonymization is not without its cost: utility and privacy of data are intrinsically connected. "Data can be either perfectly anonymous or useful but not both", says Paul Ohm. Increasing data privacy by anonymization means inevitably reducing data utility. (Ohm 2010 pp 1704–1705.) Anonymization requires a cost/benefit analysis: the benefit is the maintenance of confidentiality, while the cost is the resulting degree of information loss (Naugler 1984, p 86).

Ohm argues that anonymization of data is more difficult than it has been believed. It is true that a malicious adversary can use personally identifiable information such as a name or social security number to link data to identity, but it has turned out that the adversary can do the same thing using information that nobody would classify as personally identifiable. (Ohm 2010, p 1704). This means that "personally identifiable information" is not an immutable category. What is not personally identifiable information at one moment can be transformed into that at a later juncture. The identifiability of data depends upon context (Schwartz and Solove 2011, pp 1814, 1816, 1890.) As data gets aggregated, information that is not identifiable can be identified (Solove 2013, p 1891).

According to an extreme view, all data in the world can, in one way or another, be linked to natural persons. In that sense, all data is personal. (Szekely 2014.) There are "data fingerprints"—combinations of data shared by nobody else in the dataset—that allow at least a partial de-anonymization of data and re-identifying the people hidden in it. Once a dataset has been partially or entirely de-anonymized it can be combined with other datasets which opens further possibilities for re-identifying the people and de-anonymizing new datasets (Ohm 2010, p 1723.)

This argument and examples of successful re-identification have been criticized (Barth-Jones 2012). In the minds of computer security experts haunt super-users who are "difficult to find, immune to technological constraints, and aware of legal loopholes" (Ohm 2008). Others argue that there is no need for concern:

> "…re-identification attempts will continue to be expensive and time-consuming to conduct, require serious data management and statistical skills to execute, rarely be successful when data has been properly de-identified, and, most importantly, almost always turn out to be ultimately uncertain as to whether any purported re-identifications have actually been correct." (Barth-Jones 2016b.)

One counter-argument is that the attacker needs a perfect population register to link it to a partially anonymized dataset and to estimate how successful his re-identification has been. However, like super-users, perfect population registers are a myth. (Barth-Jones 2016a; Yakowitz and Barth-Jones 2011.) However, the UK code of anonymization takes the threat of re-identification seriously. It states that "the risk of re-identification through data linkage [of various sources] is essentially unpredictable because it can never be assessed with certainty what data is already available or what data may be released in the future." Initiatives such as open data, and the publication on the internet of information released under freedom of information legislation, mean that it is easier than ever to harvest and analyze large amounts of data. (ICO 2012, pp. 18–27.)

Information safe haven approach

The last strategy is to create an information safe haven for personal information. In this strategy, information is taken into archival custody and preserved there unmodified under access restrictions until public access is possible. This is clearly the most beneficial strategy for archives. For instance, Terry Cook (1991, p 22) finds census data essential and says that it must be preserved permanently. He adds that "census data is of course very sensitive and archivists must safeguard it from public disclosure until

such time as the sensitivity has disappeared". Elena Danielson shows by examples about the private lives of Thomas Jefferson and his de facto wife Sally Hemings, Thomas Mann, and Anne Sexton how survival of cultural heritage depends on trusted archives and well-crafted donor agreements which protect persons and their families (Danielson 2010, pp 184–185). Danielson writes

> "It is the very facts that might seem most private and embarrassing that can be essential information... It is the role of the archives to negotiate that private space and allow potentially embarrassing facts that have a wider historical significance to survive until they can be properly evaluated. The nature of taboo varies widely from one culture to another as does the perception of what is private. Certain elements are constant, and one is the need to preserve evidence from the private sphere." (Danielson 2010, pp 187–188.)

Danielson concludes that confidence in the discretion of the archives and in the enforcement of restrictions demonstrably contributes to the creation and preservation of important documentation (Danielson 2010, p 194). However, confidence is easily lost if the discretion is not used wisely. Furthermore, the society has to be willing to respect the archival institutions as sanctuaries of information. Otherwise, archives cannot have public confidence.

The same mechanism is vital in all transfers of information across spheres. Public records react to exposure like photographic film. If it is known that the information will become accessible to outsiders it starts immediately to affect the content of records. Therefore, there is the danger of a chilling effect, as has been already noted. It has been argued that "good government flourishes in the dark": only if decision-makers and civil servants can be sure that privacy of their discussions is respected, free flowing exchanges of thoughts are possible and also unpopular ideas are brought forward without fear of unpleasant repercussions (Flinn and Jones 2009, p. 50). Thus, protection of privacy is not only a problem for archives: it is also a tool for guaranteeing that full and frank documentation is generated in the first place and then preserved.

While public access to information is not possible, one can either deny all access or allow limited access. European archivists have favored lengthy blanket restrictions (75 – 120 years) which seem too long from the American point of view: they are "as long as the life of the subject, and often longer than the interest in the topic" (Danielson 2010, pp 196–200).

If limited access is allowed, there are several possible courses of action. Firstly, the archival institution can expect the donor to identify privacy concerns in the collections they transfer to the archives. Alternatively, archival institutions can shift responsibility to the users and make them responsible for the appropriate use of sensitive information. Institutions may screen researchers and control their use of information by contractual agreements. Institutional review boards by respected peers have been very successful in preventing release of damaging material. The third approach is to screen information. If screened information is only temporarily removed from the collection before access is given, the collection retains its full informational value. (Danielson 2010, pp 209–211; MacNeil 1992, pp 138–143.) Basically, these variations turn an information safe haven into a "fenced garden". Who controls contextual transfer in a fenced garden depends

on the conditions for use. It can be even the person himself who grants the access to the archives.

Summary

The choice of the strategy impacts on what information is preserved and how it can be used. Most of the strategies are paternalistic in the sense that somebody makes the decision about the fate of the information and its possible use in other contexts on the behalf of the individual. However, privacy self-management and the right to be forgotten are individualistic principles because the subject controls the use of the information. The purpose limitation principle is in this respect neutral, because it is the purpose of information collection that sets limits to the information usage. All the strategies affect or limit contextual transfer in some way (see Table 1).

Table 1 Societal strategies and contextual transfer

	Nature of the strategy	*Contextual transfer*
Purpose limitation principle	*Neutral*	*Limited (purpose-bound)*
Privacy self-management and the right to be forgotten	*Individualistic*	*Limited (by consent) or not possible*
Destruction	*Paternalistic*	*Not possible*
Anonymization	*Paternalistic*	*Yes, with loss of information*
Safe haven approach	*Paternalistic*	*Yes, with full information*

The safe haven approach is the most favorable from the archival point of view, because it allows contextual transfer without information loss. Destruction is the opposite: it prevents all contextual transfer and access to information.
The purpose limitation principle and privacy self-management together with the right to be forgotten are problematic to archives. They may allow contextual transfer, but the decision is not made by the archives and it is not based on impartial archival evaluation of information's value.

Anonymization offers the most complex set of challenges. If anonymization fails, re-identification is a potential problem when users are granted access to anonymized information. On the other hand, if anonymization is successful, archives are in danger of storing information whose value for research and society is limited or diminished from what it would have been without anonymization.

Discussion

One can easily see how societal strategies may in principle conflict with the archival mission. If information cannot be transferred to archives because of the purpose limitation principle, if it is destroyed without regard to the archival point of view, or if it is preserved but some information is lost because of anonymization, the possibilities for archives to take care of their societal function are diminished. If implementation of the strategies takes place without regard to the needs of records and archives

management—and especially cultural-historic usage of information—there is a danger that archives cannot fulfil the function that they have: transferring usable information from one point of time, place, and context to another.

However, whether there is an actual conflict at practical level depends on the concrete situation, on local legislation and how the legislation is implemented. Legislation may reconcile societal strategy with archival needs by excluding some categories of personal information from the sphere of privacy protection, by recognizing research interests or by allowing personal information in the custody of archival institutions to be disclosed (MacNeil 1992, pp 79–80; see also Ketelaar 1995). Terry Cook correctly notes:

> There are sometimes legislative or statutory prohibition which prevent archivists from viewing certain series of records in order to appraise them or which legally bar the transfer of certain categories of records to the archives, or both. In such cases, archivists (and their outside supporting communities) must lobby for legislative amendments or administrative arrangements to overcome these prohibitions. Ideally, archival legislation itself grants archivists right to appraise and acquire even sensitive records. (Cook 1991, p 60.)

Frequently, however, legal standards fail in this respect and they either do not consider the desirability of permitting research access to personal information, leave it to the discretion of government agencies, or allow access only to records in a form which is not individually identifiable (MacNeil 1992, p 90). In the case of the right to be forgotten, the data protection legal framework does not address records management considerations, or at least, not in any explicit, easily recognizable manner (Xie 2016). The Finnish recordkeeping environment offers another example of how legislation does not always lead to an optimal result. Finnish legislation often requires that personally identifiable information is destroyed when it is not needed for the specific original purpose. While legislation also often acknowledges the need to make exceptions and transfer the information to archival custody, public agencies sometimes ignore this and destroy information without considering its archival value (Arkistolaitos 2009). While destruction of information is often a necessity—also archives are forced to destroy personal information to keep the bulk manageable (see e.g. Cook 1991, p 3)—the question is not whether information should have been preserved, but who has the right to decide what information is preserved permanently and what are the grounds for this decision. If societal strategies are implemented without regard to archival needs, archival institutions have no say in this matter.

It is argued that the principle of right to be forgotten can and should only be applied in situations where the individual has consented to the processing of personal data (Ausloos 2012). This would often exclude information that is collected by public sector organizations. Also Ivan Szekely (2010) argues that the right to be forgotten has a limited applicability in archives. According to him the primary aim of historical archives is to guarantee integrity of the documents in their care whereas administrative archives aim to guarantee the truth value of the documents in their possession. The right to be forgotten has limited applicability in both cases, since the archive must not undermine the integrity of the documents, according to Szekely (2010.) However,

Szekely's argument does not show that the right to be forgotten could not be applied against archival interests.

Archives have reacted to the idea of introducing the principle of the right to be forgotten in the EU legislation with considerable anxiety, although the purpose of the right is neither to limit academic research nor to hinder the preservation of historic past (Szekely 2010). If archives is a kind of "collective" or "social memory" (Hedstrom 2010; Jacobsen, Punzalan, and Hedstrom 2013) the privacy self-management principle and the right to be forgotten suggest that a person may choose to exclude himself from this memory. The ethos of the right to be forgotten is opposed to the very idea of long term preservation of information: the right refers to a situation where a historical event should no longer be revitalized due to length of time that has elapsed since its occurrence. The longer the origin of the information goes back, the more likely personal interests prevail over public interests (Weber 2011.)

Destroying information is easily in conflict with the archival goal of preserving information. For instance, the National Archives of Finland has had disagreements with government agencies who would like to destroy information that the archives have wanted to keep. The argument of the agencies has been that the information is too sensitive to be kept in the archives (Orrman 2012.) In Finland, being hospitalized in a mental asylum was marked in census records, for instance. This information was retrospectively removed from the records in 1983. From an archival point of view the danger is that the records do not give an undistorted view of the society and the conditions in which they were created. For society, the danger is that gaining compensation for past misdoings may be difficult, if there are no records proving the harm for the individual. An example of this are the forced sterilizations that took place in Sweden in years 1935 – 1975 in part on social and eugenic grounds (Orrman 2012.) On the other hand, also archives have destroyed information to protect privacy. In the Nordic countries, this is known as "ethical appraisal". In Sweden, records about high nobility family scandals in the fonds of the Svea Court of Appeal were destroyed already in the 1860's (Landahl 1950, Orrman 2012).

The concept of privacy often leads to thinking about ways to limit access to and use of information. Privacy is a basic right in a democratic society. The right to privacy is sometimes in conflict with other rights and values and it must be balanced against them. (MacNeil 1991.) One should remember that data analysis may also have social benefits. For instance, let's imagine that over the course of the past decade a person has given out 50 000 pieces of data and has not been negatively affected by this. One day, a relative innocuous fact 50 001 gets combined with the others and reveals that the person is at risk for contracting a highly contagious and lethal disease. (Solove 2013, p 1890.)

Anonymized data has diminished value for research. Data linkage allows the research community to correlate different factors that may influence phenomena. For instance, the linkage of employment records and mortality data greatly increases the significance of research results in the area of occupational health and safety. (MacNeil 1992, pp 88–91.) In longitudinal research personal identifiers are necessary to enable the researcher to track individuals or groups over time. Similarly, correlational research which establishes relationships between characteristics of individuals and usually requires the linking of personal information held in separate records systems. (MacNeil

1992, p 134.) These possibilities are lost when data from different sources cannot be combined.

Developments in data re-identification techniques may have consequences for records and archives management practices and they should be followed closely by the records and archives management community. Paul Ohm states that the failure of anonymization forces privacy law to abandon its reliance on personally identifiable information and find an entirely new paradigm on which to regulate information privacy (Ohm 2010). Paul M. Schwarz and Daniel J. Solove instead argue that we should build two categories of personally identifiable information, "identified" and "identifiable" data, and to treat them differently (Schwartz and Solove 2011, p 1817).

The fact that personally identifiable information may be a mutable category is something that records professionals should consider in their work. Archival practices rely on the recognition of personally identifiable information in records: it is a precondition both for defining valid access restrictions and weeding personal information when necessary. Terry Cook has argued for "temporary pseudo-anonymization" of paper records stating that records containing highly personal information kept solely for their collective significance should not be made available through descriptive tools which allow retrieval by a personal identifier during the person's lifetime. (Cook 1991, pp 7, 62.) In the digital environment such approaches are likely to be less successful.

Our conception of personal information is reflected in records practices. According to Jens-Erik Mai (2016) personal information can be viewed either as a "true representations of state of affairs" or "signs which are open for interpretation and negotiation". Privacy theories that are built on the first approach focus on controlling, limiting, and restricting access to the material information carrier of information. Privacy theories in the second tradition focus on pragmatics of the information and the situation and the aim is to regulate use, analysis, and interpretation of personal information. (Mai 2016.) To my understanding, records practices today largely rest on the first approach—on controlling, limiting, and restricting access. However, considering the change of privacy issues in digital environment, archives should think more about use, analysis and interpretation of personal information. The focus should be less on defining and following access restrictions and more on questioning what can be done with the accessed data. This is in sync with the general need to have "privacy law's timing adjusted", as Daniel J. Solove has argued. Currently privacy self-management and privacy laws focus heavily on the time of the initial collection of data. However, it is often difficult to evaluate benefits and costs at that time and the decisions would be better made at the time of particular uses of data (Solove 2013, p 1902.)

The mutability of the category of personally identifiable information creates challenges for archivists when they try to reconcile different needs in the society. Archivists have the unenviable task of reconciling legitimate but conflicting interests—the individual's right to privacy and society's need for knowledge (MacNeil 1992, p 5). Archivists are aware of needs for privacy, but they have also the responsibility to promote the use of records: it is a fundamental purpose of the keeping of archives (Dever 2012). Balancing the rights of the individual with the rights of the community presents a problem to which legislative approaches offer less than ideal solutions

(MacNeil 1992, p 62). “Public interest” has no abstract definition (MacNeil 1992, p 76) and at times it is better served by secrecy than openness (MacNeil 1992, p 83). Legislation may present a greater burden on the party objecting to the disclosure of information, or vice versa on the party seeking access to personal information to prove why it should be disclosed (MacNeil 1992, p 77). Finding a middle way between various interests is not an easy task. Szekely (2010) says that satisfying the future researchers' curiosity and serving the historical knowledge of future generations cannot take place at the cost of the disproportionate infringement of the rights of the present generation. He also believes the goals can be partially attained without the infringement of the data subjects' information rights. Also Heather MacNeil (1991) asks if the research constitutes a legitimate reason for disclosing personal information. She answers by stating that “it is unlikely that the community as a whole will accept ethical standards that give higher priority to research than to respect for human subjects” (MacNeil 1991).

Conclusions

The ways privacy is protected in the society are significant for archives. They impact the documentation that is generated in society, the extent to which this documentation is preserved, and what kind of processes lead to decisions about preservation of information. Records and archives management profession should be aware of possible conflicts between the societal strategies and the archival mission and to be ready to defend their perspective to get the archives recognition and acceptance in society. The question is more than “archival” because it touches all the contexts that the records may have. Records professionals should approach the issue holistically embedding the management of privacy by design in recordkeeping, as ISO 15489-1:2016 suggests.

Digitalization has changed privacy issues. Reconciling different needs is becoming harder than before because identifying personally identifiable data may be more difficult to recognize. One should pay more attention than before to the possibilities of using the information instead of only controlling access to it. This creates pressure to develop new methods for privacy management that are less dependent on the early identification of personal data and definition of fixed access restrictions and more focused on the examination of possibilities to using the data. The first European code of practice on anonymization (ICO 2012, p. 4,) provides a starting point for that. Ideally, there would be a methodology for assessing risks that may result from linking the data with other perhaps in part unknown sources.

References

An X (2003) An integrated approach to records management. The Inf Management J, 37(4): 24–30

Arkistolaitos (2009) Selvitys henkilörekistereihin sisältyvien tietojen säilyttämisestä ja poistamisesta [Report about destruction and preservation of personal data]. Dnro AL/17052/07.01.01.03.00/2009

Ausloos J (2012) The “right to be forgotten” - Worth remembering? Computer Law

and Security Rev 28(2): 143–152. doi: 10.1016/j.clsr.2012.01.006
Barth-Jones D (2012) The “re-identification” of Governer William Weld’s medical information. A critical re-examination of health data identification risks and privacy protections, then an now. doi: 10.2139/ssrn.2076397
Barth-Jones D (2016a) Re-identification risks and myths, superusers and super stories (part I: risks and myths). https://concurringopinions.com/archives/2012/09/re-identification-risks-and-myths-superusers-and-super-stories-part-i-risks-and-myths.html. Accessed 21 Oct. 2016
Barth-Jones D (2016b) Re-identification risks and myths, superusers and super stories (part II: superusers and super stories). https://concurringopinions.com/archives/2012/09/re-identification-risks-and-myths-superusers-and-super-stories-part-ii-superusers-and-super-stories.html. Accessed 21 Oct. 2016
Bingo S (2011) Of provenance and privacy. Using contextual integrity to define third-party privacy. The American Archivist 74(Fall/Winter): 506–521
Brennen S, Kreiss D (2014) Digitalization and digitization. http://culturedigitally.org/2014/09/digitalization-and-digitization. Accessed 23 May 2017
Cook T (1991) The archival appraisal of records containing personal information: a RAMP study with guidelines. Unesco, Paris
Danielson ES (2010) The Ethical Archivist. Society of American Archivists. Chicago
Dever M (2012) The private in the public archive. In: Carucci M (ed) Revealing privacy. Debating the understandings of privacy. Peter Lang GmbH, Internationaler Verlag der Wissenschaften, Frankfurt am Main, pp 19–29
Duchein M (1983) Obstacles to the access, use and transfer of information from archives: a RAMP study. Unesco, Paris
Flinn A, Jones H (2009) The Freedom of Information Act in practice: the historian’s perspective. In: Flinn A, Jones H (eds) Freedom of Information. Open Access, Empty Archives? Routledge, London and New York, pp 33–53
Gellman R (2014) Willis Ware’s lasting contribution to privacy: Fair information practices. IEEE Security and Priv 12(4): 51–54. doi: 10.1109/MSP2014.82
Gellman R (2016) Fair information practices. A basic history. Version 2.16. http://bobgellman.com/rg-docs/rg-FIPShistory.pdf. Accessed 23 Sept. 2016
Greenleaf G (2012) The influence of European data privacy standards outside Europe. Implications for localisation of Convention 108. International Data Privacy Law 2(2). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1960299. Accessed 7 Oct. 2016
Hedstrom M (1981) Computers, privacy, and research access to confidential information. The Midwest Archivist 6(1): 5–18
Hedstrom M (2000) How to Proceed? Presenting the results of a working meeting on Recordkeeping Metadata. Summary. http://www.archiefschool.nl/docs/hedshowt.pdf %5B24.10.2001%5D. Accessed 24 Nov. 2001
Hedstrom M (2001) Recordkeeping metadata. Presenting the results of a working

meeting. Archival Science 1 243–251. doi: 10.1023/A:1012483201735
Hedstrom M (2010) Archives and collective memory. More than a metaphor, less than an analogy. In: Eastwood T, MacNeil H (eds) Currents of archival thinking. ABC Clio, Santa Barbara, CA, pp 163–180
Hubaux JP, Juels A (2016) Privacy is dead, long live privacy. Protecting social norms as confidentiality wanes. Communications of the ACM 59(6): 39–41. doi: 10.1145/2834114
Hull G (2015) Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data. Ethics and Inf Technology 17(2): 89–101. doi: 10.1007/s10676-015-9363-z
Huskamp Peterson T (2007) Privacy is not a rose. http://trudypeterson.com/wp-content/uploads/2014/documents/Privacyisnotaroserev.doc. Accessed 3 Aug. 2016
ICO (2012) Anonymisation: managing data protection risk. Code of practice. Wilmslow, Information Commisioner's Office. Retrieved from http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf Accessed 5 June 2017
ISO 15489-1 (2016) Information and documentation – Records management. Part 1 Concepts and principles
Jacobsen T, Punzalan RL, Hedstrom, ML (2013) Invoking "collective memory": mapping the emergence of a concept in archival science. Archival Science 13: 217–251
Ketelaar E (1995) The right to know, the right to forget? Personal information in public archives. Archives and Manuscripts 23(1): 8–17
Koops BJ (2011) Forgetting footprints, shunning shadows. A critical analysis of the "right to be forgotten" in big data practice. SCRIPTed 8(3). doi: 10.2966/scrip 080311.229
Landahl S (1950) Skoglar Bergström och gallringen i Svea Hovrätts arkiv [Skoglar Bergstrom and appraisal of the archives of the Svea Court of Appeal]. In: Donum Boëtianum. Arkivvetenskapliga bidrag tilläagnade Bertil Boëtius 31/1/1950. Stockholm
Van Maanen J, Pentland TB (1994) Cops and auditors: the rhetoric of records. Sage Publications, Thousand Oaks, London, Delhi
MacNeil H (1991) Defining the limits of freedom of inquiry. The ethics of disclosing personal information held in government archives. Archivaria 32(Summer): 138–144.
MacNeil H (1992) Without consent. The ethics of disclosing personal information in public archives. The Society of American Archivists and Scarecrow Press, Lanham Maryland and London
Mai JE (2016) Personal information as communicative acts. Ethics and Inf Technology 18(1): 51–57. doi: 10.1007/s10676-016-9390-4
Mayer-Schönberger V (2011) Delete. The virtue of forgetting in the digital age. Princeton University Press, Princeton and Oxford

Naugler H (1984) The Archival appraisal of machine-readable records: a RAMP study with guidelines. Unesco, Paris

Nissenbaum H (1998) Protecting privacy in an information age: the problem of privacy in public. Law and Philosophy 17: 559–596. doi: 10.2307/3505189

Ohm P (2008) The myth of the superuser. Fear, risk, and harm online. UC Davis Law Rev 41(4): 1327–1402

Ohm P (2010) Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev 57: 1701–1777

Orrman E (2012) Asiakirjojen seulonnasta ja julkisuudesta tutkimuksen edellytyksiä säätelevinä tekijöinä [Archival appraisal and access to documents as factors defining conditions of research]. Tieteessä tapahtuu 3: 26–33

Pasquier M, Villeneuve JP (2007) Organizational barriers to transparency: a typology and analysis of organizational behaviour tending to prevent or restrict access to information. International Rev of Administrative Sciences 73(1): 147–162. doi: 10.1177/0020852307075701

Rauhofer J (2014) “Look to yourselves, that we lose not those things which we have wrought.” What do the proposed changes to the purpose limitation principle mean for public bodies’ rights to access third-party data? International Rev of Law, Computers and Technology 28(2): 144–158. doi: 10.1080/13600869.2013.801592

Schonsheck J (1997) Privacy and discrete “social spheres.” Ethics and Behav 7(3): 221–228

Schwartz PM, Solove DJ (2011) The PII problem. Privacy and a new concept of personally identifiable information. New York University Law Rev 86: 1814–1894

Shepherd E (2015) Freedom of information, right to access information, open data: who is at the table? The Round Table 104(6): 715–726. doi: 10.1080/00358533.2015.1112101

Solove DJ (2013) Introduction: privacy self-management and the consent dilemma. Harvard Law Rev 126: 1880–1903

Solove DJ, Scott RE (2006) A taxonomy of privacy. University of Pennsylvania Law Rev 154(3): 477–564

Szekely I (2010) The four paradigms of archival history. J of Inf Technology Res 3(4): 51–82

Szekely I (2014) The right to be forgotten and the new archival paradigm. In: Ghezzi A. Pereira ÂG, Vesnić-Alujević L (eds) The ethics of memory in a digital age. Interrogating the right to be forgotten. Palgrave Macmillan UK, London, pp 28–49. doi: 10.1057/9781137428455_3

United Nations (1948) Universal declaration of human rights. http://www.un.org/en/universal-declaration-human-rights. Accessed 28 June 2017

Vistilä M, Ruokonen F (2016). Social networking sites and privacy as contextual integrity. In: Carucci M (ed.) Revealing privacy. Debating the understandings of privacy. Peter Lang GmbH, Internationaler Verlag der Wissenschaften, Frankfurt

am Main, pp 119–132
Weber RH (2011) The right to be forgotten: more than a Pandora's box. JIPITEC 2(2): 120–130
Worthy B (2010) More open but not more trusted? The effect of the Freedom of Information Act 2000 on the United Kingdom central government. Governance 23(4): 561–582. doi: 10.1111/j.1468-0491.2010.01498.x
Xie SL (2016) Retention in "the Right to be forgotten" scenario. A records management examination. Records Management Journal 26(3): 279–292. doi: 10.1108/RMJ-11-2015-0038
Yakowitz J, Barth-Jones D (2011) The illusory privacy problem in Sorrell v . IMS Health. Technology Policy Institute (May): 1–8. http://www.ehealthinformation.ca/wp-content/uploads/2014/08/the-illusory-privacy-problem-in-sorrell1.pdf
Young JB (1978). Privacy. Wiley, New York