Threats of the data-flood

Jeurgens, K.J.P.F.M.

Published in:
Archives in Liquid Times

CHARLES JEURGENS

Threats of the data-flood. An accountability perspective in the era of ubiquitous computing.[1]

Overview

In this essay, I argue that ubiquitous computing and the closely related increase in data requires a fundamental reorientation of the recordkeeping community. I explore the effects of data-driven phenomena like big data and smart applications on records and recordkeeping practices from the perspective of its contribution to informational accountability and transparency. I contend that a traditional view of appraisal of recorded data is no longer sufficient to contribute to accountability and transparency. Instead, the focus should be shifted to understanding and managing the assemblages between data and the processing mechanisms (for instance algorithms) in situated practices.

There would indeed be no archive desire without the radical finitude, without the possibility of forgetfulness which does not limit itself to repression.
Jacques Derrida

Introduction

In the mid 1970s, the Italian writer Italo Calvino masterfully depicts the ritual of emptying the trash. In his tale, *La poubelle agréée* he demonstrates the struggle between retaining and discarding. The way people treat their waste reflects the essence of being human, or as Calvino states: "[a]las the unhappy retentive (or the miser) who, fearing to lose something of his own, is unable to separate himself from anything, hoards his faeces and ends up identifying with his own detritus and losing himself in it' (Calvino, 1993, p. 58). Calvino's main character is in a persistent quandary about how to distinguish between the essential and the residue, the meaningful and the meaningless, the relevant and the extraneous. But the perception of what is waste and what is valuable has changed fundamentally in the last few decades. One of the largest European sanitation companies now advertises with the slogan 'waste doesn't exist', since everything can be recycled and reused in the circular economy. This changing perspective bears strong resemblance with one of the core functions the recordkeeping profession is traditionally engaged with: managing abundance by identifying records to be curated and preserved and what

[1] I would like to thank Geert-Jan van Bussel, Annet Dekker and Eric Ketelaar for their comments on an earlier version of this article.

should be discarded. But, analogous to the world of sanitation, the dividing line between valuable information and worthless trash is rapidly blurring.
The recordkeeping community is confronted with this new dilemma since the pervasive recording of data creates unprecedented opportunities in many different domains like health care, crime fighting and societal convenience in smart applications.

Data driven phenomena like Big Data, smart cities and the Internet of things are widely seen as heralds of fundamental societal transformation in a world in which everyone and everything is always connected via information networks. The implications of the computational turn go far beyond the instrumental use of ICT. More fundamental is that the world is increasingly interpreted and explained in terms of data and information. Dutch philosopher Jos de Mul calls it the 'informatisation of our worldview' (De Mul, 2002, p. 130-134). Luciano Floridi designates this turn as the fourth revolution (the three preceding were based on the observations and new paradigms of Copernicus, Darwin and Freud) since human agency in society is entirely determined by ICT which surrounds us. The effect of this informational revolution is, like the previous ones, a fundamental rethinking and repositioning of ourselves into the world (Floridi, 2014, p. 87-94).

The desire to track and monitor nearly everything is not new. States are infamous collectors of information; it is even a prerequisite to possess enough information to be able to create a political space. In 1840 the French politician Pierre-Joseph Proudhon asserted that '[t]o be ruled is to be kept an eye on, inspected, spied on, regulated, indoctrinated, sermonized, listed and checked off, estimated, appraised, censured, ordered about. (…) To be ruled is at every operation, transaction, movement, to be noted, registered, counted, priced, admonished, prevented, reformed, redressed, corrected' (quoted by Scott, 1998, p. 183). What is new are the information and communication technologies that 'record, transmit and, above all, *process* data, increasingly autonomously' and the effect is a strong belief that by doing so society will improve (safer and better quality of life) (Floridi, 2015, p. 52). According to some scholars this makes it viable for governments to record almost everything what people do or say (Villasenor, 2011). Big Data adherents are convinced of the value of data per se and they challenge the necessity of managing information based on the principles of the past. Ralph Losey, an active eDiscovery lawyer foresees that the traditionalist information-management approach based on 'classification, retention, and destruction of information' will be completely superseded within five years. In his view, the 'classify and control lock-down approach of records-management is contrary to the time. Instead of classify and kill, [it is] the *googlesque approach of save and search*'.[2] According to these data hoarders the real efforts to be made are directed towards refining methods of identifying relevant information. Keeping data will become default because as Clay Shirky stated: the problem is 'not information overload. It's filter failure'.[3] The idea of

[2] https://e-discoveryteam.com/2015/02/08/information-governance-v-search-the-battle-lines-are-redrawn/?blogsub=confirming#blog_subscription-3 accessed 30 March 2017.

[3] https://www.youtube.com/watch?v=LabqeJEOQyI, accessed 30 March 2017. His vision was disputed by Nicholas Carr, who responded 'It's not information overload. It's filter success' which means that filters push growing amounts of information that is of immediate interest to us, with the result of increasing information overload for individuals, available at <http://www.roughtype.com/?p=1464> accessed at 30 March 2017.

keeping all data, however, is not undisputed. Many scholars envision the future of information overload in terms of getting stuck in a meaningless data swamp. Jennifer Gabrys sketches the danger of transforming the archives into sites of digital rubbish because '[t]he transience and even banality that emerge with electronic storage extends to new levels, where heartbeats and expiring milk acquire a place as archive-worthy data. In fact, through the monumental task of archiving everything, the archive becomes more akin to a disorderly waste site, which then requires processes of computation to make sense of the welter of material and data' (Gabrys, 2011, p. 120).

In this essay, I explore the implications of this fourth revolution for archival memory functions in society and more specifically to understand what effects these data-driven phenomena have on the traditional function of appraisal with regard to accountability. I will argue that the recordkeeping community needs to put more effort in rethinking and redefining the prevailing archival concepts and archival functions. I contend that appraisal remains a meaningful activity in this 'age of zettabyte' (Floridi, 2014, p. 13), but that the perspective of appraisal in twenty-first century informational practices is no longer confined to reducing the volume of records but expanded with the question which components of the constructing layer of the record are required to keep the quality of records as instruments of accountability.

Radical turbulences

New technologies that generate, store and transmit data, are changing the nature of the archive. Geoffrey Batchen writes that the 'archive is no longer a matter of discrete objects (files, books, art works etc) stored and retrieved in specific places (…). Now it is also a continuous stream of data, without geography or container, continuously transmitted and therefore without temporal restriction (…)' (Batchen, 1998, p. 49; Batchen, 2001, p. 183). The change is not only related to the abundance of data. Derrida emphasised the importance of understanding the implications of technologies of communication and recording for the archive. He coined the term archivisation to express the pivotal impact of the technical means and methods on what can be archived: 'the technical structure of the *archiving* archive also determines the structure of the archivable content even in its very coming into existence.' The performative implications of that notion are far-reaching, since 'archivization produces as much as it records the event' (Derrida, 1998, p. 17; Manoff 2004, p. 12). In his *Mal d'archive*, which was published in 1995, Derrida envisaged how for example email will transform the entire public and private space since '[i]t is not only a technique, in the ordinary and limited sense of the term: at an unprecedented rhythm, in quasi-instantaneous fashion, this instrumental possibility of production, of printing, of conservation, and of destruction of the archive must inevitably be accompanied by juridical and thus political transformations' (p. 17). The adoption of email in the 1990s is an example of what Derrida called 'radical and interminable turbulences' (p. 18). New media transform what can be recorded and archived, and thus what can be used as evidence. The invention of the camera and phonograph in the nineteenth century are well known examples of the past. In our time, technologies of Big Data and Internet of Things cause unprecedented interminable turbulences. In the next

paragraphs, I will first explore the transformative effects of these technologies, then give some examples and I will finish with discussing the implications for recordkeeping concepts and for appraisal and selection.

A need to rethink archival methods

Some leading archival scholars like Frank Upward and Barbara Reed argue that the archives and record profession is facing a widespread crisis. One of the obvious signs of being in crisis is that professionals cannot 'reliably say what a record as a thing is as our conceptual understanding of it blurs into data, documents, information, the archive, and the plurality of archives. The settings in which we manage these converged "things" continues to multiply and increase in complexity. Our new information spaces with their vibrant diversity are paradoxically producing a collapse of collective memory' (Upward et all, 2013, p. 40). There are some parallels to be made with the alarmist view David Bearman already expressed in the late 1980s, when he proclaimed that 'the best methods of the profession were inadequate to the task at hand' (Bearman, 1989, preface). Since Bearman vented his concern, the information-scape has been constantly in transformation. In his time, the late 1980s, the administrative use of Internet was still in its infancy. Tim Berners Lee had just started to work on what would become the world-wide web. Social media were not born yet and the first sms would be sent in 1991. Big Data and the Internet of Things were still a science fiction fantasy. Most of these new media are commonly used nowadays. The computational turn not only affected information and communication behaviour in the personal realm but it profoundly transformed information and communication patterns in administration and business. The computational turn enabled the rise of new economic models which are based on sharing commodities and services, with Airbnb and Uber as the best-known examples. Despite the major changes in the use of ICT, the debate on appraisal and selection has largely remained within the existing document-oriented paradigm. Recently, the Australian Recordkeeping Roundtable paid attention to the implications of the computational turn on recordkeeping functions, including appraisal and selection. Kate Cumming and Anne Picot presented a valuable overview of the challenges appraisal and selection are confronted with. Some of them were diagnosed as technical (new media and applications, networks, changing forms of records, data volumes and storage) and others as organisational (multiple professional responsibilities, decentralised business processes, commercialisation and proprietary systems) (Cumming & Picot, 2014, p. 133-145). They conclude that appraisal in archival institutions is still too much defined as 'a process to preserve a documentary cultural heritage rather than identifying appraisal as laying the basis for practical and accountable recordkeeping'. Although the authors delineate some valuable directions that need to be explored to rethink and reformulate appraisal and call for developing a strategy to prioritise and to employ with business operations, they pay relatively little attention to the fundamental changes that digitisation and informatisation of society have on the attributed function(s) of appraisal. This brings up the following question: what is needed for 'accountable recordkeeping'?

Ubiquitous information technology

In 2011 the authoritative Dutch Scientific Council for Government Policy warned against a precarious lack of awareness among policy-making officials about the far-reaching implications of the networked information structures for the memory functions of iGovernment. The Council emphasised that '[b]oth the importance of 'forgetting' – people should not be judged eternally on the information that government has stored about them – and of saving and archiving require a radical cultural transition and a firmly grounded strategy' (WRR 2011, p. 16 and p. 207). The Council asserted that the government has changed from eGovernment – in which ICT is mainly directed towards providing services – into iGovernment – where ICT changes the relationship between government and citizens because information-flows and data-networks are used for purposes of control and care. The ubiquitous use of memory chips in innumerable applications and functions leads to unprecedented volumes of recorded and processed data. Beyond the three V's (the availability of high volumes, high velocity and high variety of data), it is especially the ability to search, aggregate, and cross-reference large data sets that generate these unprecedented opportunities (Boyd & Crawford, 2012, p. 663). As a result of these innovations, Chris Anderson, editor-in-chief of WIRED magazine, announced the death of theory in 2008 in his much-discussed, contested but nonetheless influential article in *Science* by stating: '(…) faced with massive data, this approach to science – hypothesize, model, test – is becoming obsolete. (…) There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot' (Anderson, 2008). Computer scientist Jim Gray introduced the fourth paradigm of science in 2007. After empiricism (observation and experiment), theory (using models, generalisations, hypotheses) and computation (simulating complex phenomena), science is increasingly based on data intensive computing, which unifies theory, experiment and simulation (Hey cs, 2009). This mixing up of correlation and causality and this naïve belief in the power and possibilities of data to solve present-day problems is typical for these big data adherents.

Data in itself might be seen as innocent, but the processing is definitely not (Rouvroy & Berns, 2013). It is the processing activity that makes data meaningful and transforms data into information. Transforming data into meaningful information cannot exist without a selective perspective. The terms data and information are often improperly used as synonyms. Liebenau and Backhouse make a clear distinction between data and information by defining data as 'symbolic surrogates which are generally agreed upon to represent people, objects, events and concepts' while information is 'the result of modelling, formatting, organising, or converting data in a way that increases the level of knowledge for its recipient', or as they summarise: 'information is data arranged in a meaningful way for some perceived purpose' (Canhoto & Backhouse, 2008, p. 48). The techniques used for modelling and organising data are increasingly computational algorithms. Algorithms are basically a set of rules or instructions to perform a certain assignment in order to process input into output. In the words of the Norwegian media scholar Eivind Røssaak, computational algorithms have become the new lingua franca of codes in the informational infrastructure and they increasingly

rule society and our lives (Røssaak, 2016, p. 34). Compared to human processing, computational algorithms have many advantages since they are much faster, can deal with more complexity and are more accurate than humans will ever be. The downside of this computational processing is that these systems rely on processes and abilities 'that are radically beyond what is possible for human beings to understand' (Danaher, 2016, p. 247). They are black boxes and that is what gives rise to many concerns, because we do not understand how these algorithms operate as the new power brokers in society (Diakopoulos, 2014, p. 2). Critics like Evgeny Morozov and Cathy O'Neil stress that algorithms are constructed models, based on choices what to include and what to leave out. And these choices 'are not just about logistics, profits and efficiency. They are fundamental moral' (O'Neil, 2016, p. 218; Morozov, 2014, p. 182-186). John Danaher warns that the increasingly reliance on algorithms in decision making processes might turn society in an 'algocracy', a governance system in which computer-programmed algorithms are used 'to collect, collate and organise the data upon which decisions are typically made, and to assist in how data is processed and communicated through the relevant governance system' (Danaher, 2016, p. 247). While in a bureaucracy laws and regulations structure and enforce how humans act, in an 'algocracy' the algorithms are the structuring and constraining components. Janssen and Kuk emphasise that algorithms do not work on their own, but form an 'algorithmic materiality', which means that there is an intricate relationality between algorithm, systems, data and humans resulting in a dynamic and 'complex socio-technical ensemble of people, technologies, code developers and designers' (Janssen & Kuk, 2016, p. 274-275), which is very similar to the archive as 'the apparatus through which we map the everyday' (Giannachi, 2016, p. xv).

A few societal examples discussed

There are good reasons to worry about this emerging 'algorithmic governmentality' as some scholars label this data-driven exercise of power and policymaking (Thomas & Berns, 2013; Rodrigues, 2016). Before discussing the archival implications of these socio-technical developments, I want to review some examples. Real time processing of large quantities of data from criminal records, police databases and surveillance data to predict where criminal activities are likely to happen (predictive policing) has already been put into practice in several countries (Joh, 2015). Even in the courtroom computational algorithmic support has been introduced to underpin court decisions. The independent non-profit organisation of investigative journalism *ProPublica*, recently published a series of critical articles on the accurateness of algorithms used in courtrooms to assess the likelihood of recidivism of defendants. *ProPublica* journalists analysed the accuracy of a widely-used risk assessment tool named COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) by investigating 10,000 criminal defendants in Florida and compared their predicted recidivism rates with the actual rates. The researchers found out 'that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk'.[4] Although the Supreme Court of Wisconsin expressed its concern about this race correlation in COMPAS, in an appeal from an order of a circuit court it judged that the evidence-based risk assessment tool COMPAS can be used at sentencing.[5] In the explanation

of its decision, the Supreme Court circumscribed its use by stressing that risk scores are 'not intended to determine the severity of the sentence or whether an offender is incarcerated' and that risk scores 'may not be considered as the determinative factor in deciding whether the offender can be supervised safely and effectively in the community'. Although the Supreme Court agreed that the defendant-appellant was not able to review and challenge how the COMPAS algorithm – which is part of the trade secret of the developer Northpointe Inc. – calculates risk, the order judged the ability to review and challenge the resulting risk scores as satisfactory.[6] Interestingly, one of the judges, Shirley S. Abrahamson wrote a separate consideration in which she emphasised the relevance of recording the use of risk assessment tools. Precisely because scholars were critical on using these risk assessment tools in sentencing, courts should 'evaluate on the records the strengths, weaknesses, and relevance to the individualized sentence being rendered of the evidence based tool (or, more precisely, the research-based or data-based tool)'. Abrahamson recognised that this might be an extra demand on and administrative burden for the circuit courts, 'but making a record, including a record explaining consideration of the evidence based tools and the limitations and strengths thereof, is part of the long-standing basic requirement that a circuit court explain its exercise of discretion at sentencing'.[7]

We need to question how the record is defined if the judges accept that algorithms can be used in sentencing although the algorithm itself, the lens through which the data are filtered, sorted etc., remains a black box because of the mentioned trade secret. The record-making as defined by the Supreme Court has to do with accountability of how the judges use the tools in the process of sentencing, not with the processing activity of the algorithms themselves. This appeal clearly shows the limitations of the traditional scope of the concept of the record. If the informational algorithm remains a closed black-box in cases with far-reaching consequences for citizens (even if the outcomes can only be used as additional information for decisions) the claim that records provide the best means for warranting accountability is severely affected. The ever-increasing interrelationship between man and technology requires a clearer notion of the scope of the record, especially when a relation is made to accountability of decision-making. There are good reasons to redefine the scope of the record in that tight relationship between humans and machines. I agree with Amelia Acker, who argues that examining the infrastructure of records, 'archivists can think big enough about the "black box" and all the layers of construction behind digital records and emerging documentation practices' (Acker, 2016, p. 294-295). One of these layers of construction are the algorithms that are used in the processing of data. The use of 'black-box' algorithms in decision-making processes will be mirrored in the records that are created, and it is not without consequence to the attributed quality of the records as means of accountability. It is imaginable that for some decision-making processes (which immediately shows an additional selection perspective) open and understandable algorithms are required. That is for instance the motive of a motion for the European

[4] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm accessed at 30 March 2017

[5] Supreme Court of Wisconsin, State of Wisconsin versus Eric L. Loomis on certification from the court of appeals, 13 July 2016, available at <https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690> accessed at 30 March 2017.

[6] Ibid., par. 53.

[7] Ibid., par. 141

Parliament Resolution on Robotics, debated in 2017. Article 12 of this resolution says that it should always be possible to supply the rationale behind any decision taken with the aid of Artificial Intelligence (AI) that can have a substantive impact on one or more persons' lives. It must always be possible to reduce the AI system´s computations to a form comprehensible by humans. Interestingly that same article articulates the necessity that 'advanced robots should be equipped with a 'black box' which records data on every transaction carried out by the machine, *including* (my italics CJ) the logic that contributed to its decisions'.[8] This is in line with the regulation *on the protection of natural persons with regard to the processing of personal data and on the free movement of such data* which was adopted in April 2016 by the European Union. This regulation sets rules and requirements for data-driven automated processing. Every person has 'the right not to be subject to a decision (...) which is based solely on automated processing (...)'.[9] Examples that are explicitly mentioned are automatic refusal of an online credit application and e-recruiting practices without human intervention. Predictive profiling, one of the most fundamental intrusions in a private life which is increasingly used by authorities in fighting crime and terror remains allowed 'where expressly authorised by Union or Member State law'.

Big data analysis is useful for revealing general patterns, but, and that is not always kept in mind, there is always a probability of a mismatch between general patterns and a specific situation (WRR, 2016, p. 27). Scholars, journalists, advisory and legislative bodies warn against excessive techno-dependency and techno-optimism.

The Dutch investigative journalist Dimitri Tokmetzis criticises the naïve way rules are formulated and used in algorithms without paying enough attention to the validity of the underlying assumptions. To give an example: Dutch government assumes that poverty is a risk factor for the education of children. It is possible to design an algorithm to find evidence of poverty in the electronic child records and to make a list of families that need to be watched closely to be able to intervene if necessary. But do we know whether the assumption behind the rule is valid? Is the assumption that defines the rule based on thorough scientific research? (Tokmetzis, 2012, p. 59-60). Scott Mason, researcher at Keele University, also warns against the often-careless way how bureaucrats and policy-makers interpret and contextualise the results of Big Data itself without consulting domain experts to assess the validity of correlations. He is very critical about the claim that Big Data analysis creates the possibility for 'neutral' evidence based policy-making. According to Mason, 'the vast quantities of correlations generated by Big Data analytics act simply to broaden the range of 'evidence from which politicians can chose to support their arguments' (Mason, 2016). In 2016, the Dutch Scientific Council for Government Policy notified a highly undesirable tendency of policymakers to accept the revealed patterns without questioning the validity of the results for specific situations (WRR, 2016).

Archival implications

The aforementioned examples show that computational algorithms increasingly become an integrated part of government processes and decision making. Legal scholars have argued for more than 20 years in favor of more transparency in

automated processing (Kroll, 2015, p. 6). Since records which are created in the course of business, 'provide evidence of actions, decisions, and intentions, both legal and illegal, proper and improper, and wise and misguided' (Cox & Wallace, 2002, p. 4), availability of records is vital for accountability and transparency. Also in the recordkeeping realm, algorithmic tooling is used to manage the growing number of documents and to find relevant information for a specific purpose. The most advanced developments of algorithmic computation in the recordkeeping sphere can be found in eDiscovery and information retrieval applications.[10]

Since archivists claim to play a pivotal role in defending institutional and societal transparency and accountability (Jimerson, 2009, p. 246-252), there is an urgent need for archivists to ruminate what it means to take this role in the era of ubiquitous computing. If records, archives and archivists want to continue to be key players in ensuring and defending accountability, this evokes the question what meaningful recordkeeping is in this new context of data-ubiquity, and at the same time what meaningful records are.

As Upward and others have put forward, this is exactly one of the main challenges the archival and recordkeeping community is confronted with: to clarify how the conceptual relationship between data, records and archives is designated in the era of ubiquitous computing. The traditional record was based on fixity and stability in a material sense. What has fundamentally changed is the possibility to produce different aggregates out of the same recorded data, which means, as Bruno Latour (2009) writes, 'that the whole has lost its privileged status' which makes us aware of the fact that the whole is always simpler than the parts (p. 198). The written record used to have the shape of an entity (the whole) in which the parts (words, sentences, paper, lay out, signature etc.) were a fixed materialised aggregate. Since the computational turn, it is possible to use the same recorded parts (data) in different configurations simultaneously. The stable whole has been replaced by a 'continually evolving liquid assemblage of action' (Introna, 2016, p. 19). It is as if we construct different types of houses with the same bricks at the same time.

This (informational) fluidity is an important feature of what was designated by Deleuze as an assemblage. An assemblage is in the words of Deleuze 'a multiplicity which is made up of many heterogeneous terms and which establishes liaisons, relations between them (...). [T]he assemblage's only unity is that of co-functioning' (Deleuze & Parnet, 2007, p. 69). In an assemblage, an element can be dissociated from a specific assemblage and continue to function in another assemblage. Assemblages exist merely because of the relationships between the elements. Deleuze emphasises that an assemblage is never technological: '[t]ools always presuppose a machine, and the machine is always social before being technical.

---

[8] European Parliament, Motion for a European Parliament resolution. Report with recommendations to the Commission on Civil Law Rules on Robotics, A8-0005/2017, art. 12, available at <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN> accessed at 30 March 2017.

[9] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data*, and repealing Directive 95/46/EC (General Data Protection Regulation), art 71, available at <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN> accessed 30 March 2017.

[10] The Information Governance Initiative Community provides an interesting overview of activities in these fields: http://iginitiative.com/community/

There is always a social machine which selects or assigns the technical elements used' (Deleuze & Parnet, 2007, p. 70). This is an important notion which might be helpful to disentangle the sometimes-confusing relationship between data and records.

In a world of ubiquitous computing the ability to define data-points and to monitor and record data has become infinite. CISCO expects that in 2020 more than 50 billion devices are connected to the Internet and these devices 'require minimal human intervention to generate, exchange and consume data' (Rose cs, 2015, p. 17). In the past, it was a time-consuming human activity to select the elements worthwhile to be recorded. That process, the 'conscious or unconscious choice (determined by social and cultural factors) to consider something worth archiving' was coined by Eric Ketelaar as archivalisation (Ketelaar, 1999). Archivalisation precedes archiving and to understand this process, we need to understand what Hofstede called the 'software of the mind' which is programmed by social and cultural factors and comes very close to Deleuze's 'social machine'. Nowadays all particles that are 'observed' by a machine are recorded, although, and that is not unimportant, the data points to be monitored and recorded still need to be defined and programmed. The recorded raw data in itself is meaningless; these are signals without real significance. Data are only meaningful in relationship with other data, processed in a specific situation. Only from that perspective the concept of the record or archive is a meaningful construct. It means that archives should be seen as Foucauldian apparatuses (of governance), dispositives, machineries of seeing, but, and that has to be emphasised, machineries of seeing from a particular point of view (Giannachi, 2016, p. xv-xvii; McQuillan, 2016, p. 8). Thinking about the archive via the apparatus means focusing on the networked arrangement of media, mechanisms of communication and data processing (Packer, 2010). The archive is meaningless without understanding the interdependency of these socio-technical components and humans. Only then we will be able to understand, as Geoffrey Bowker writes, that every act of permitting 'data into the archive is simultaneously an act of occluding other ways of being, other realities. The archive cannot in principle contain the world in small; its very finitude means that most slices of reality are *not* represented' (Bowker, 2014, p. 1797, italics CJ). One could argue that, compared to the past practices of recording, the number of potential witnesses within a situated practice have incredibly increased by the explosion of sensors and data-points. Nevertheless, what is represented by records is in the end based on situated needs, that define the technical arrangements.

Back to appraisal

I opened this article with Calvino's quandary how to distinguish between the meaningful and meaningless. I conclude with the proposition that in our time of ongoing datafication of society, archivists need to redefine the record and as a consequence of it the recordkeeping mechanism of distinguishing between the essential and residue. I showed that protecting informational accountability requires rethinking the components of the record or archive. The apparatus-view, in which the archive functions as a machine of governance, is helpful to understand the intricate, assemblage-based relationality between the components of the archive. The recorded data is only one element of that machine. What data is relevant, and

which other components are required to create a meaningful record, is defined by the situated context of operation. The Volkswagen emission scandal of 2015 may serve as an example. Advanced software in the diesel engines of Volkswagen could detect when the car was tested and subsequently adapt its emissions during the artificial test circumstances to acceptable levels. Back on the road, the vehicles switched to normal mode with much higher emission rates. It is an example that shows that the recorded data of the tests can only be understood in combination with the software, and that the software can only be understood if the logic of the algorithms is known. If the record only provides the recorded data, it is not sufficient for informational accountability. Above all, algorithms are models written with a specific purpose and they are not neutral nor objective. Understanding the results of algorithmic processing requires at least knowledge of the underlying assumptions of the model and the data which are used by the algorithms. The familiar principle of 'the context is all' is also applicable in this layer of construction of the record.

This has major implications for the issue of appraisal, which gets a much wider scope than just answering the question of keeping or discarding recorded data. From a recordkeeping perspective, the issue is not about data; it is about what people, institutions and communities want to be able to reconstruct for purposes of business, evidence, accountability and memory. That perspective is decisive for answering the question which components of 'the archive as an apparatus' should be preserved in coherence. Providing robust accountability is not an easy and especially not a pure technical task to accomplish. Joshua Kroll, who developed a general framework for accountable algorithms in automated decision-making processes, stresses that accountability requires the possibility to verify that 'the social, legal and political structures that govern an automated process function as they are intended to function' (Kroll, 2015, p. 210). Robust accountability requires involvement in the system design and computer systems should be designed in a way that they are reviewable (Kroll, 2015, p. 188-202). Cathy O'Neil started a business to audit algorithms. In an interview with the Los Angeles Times, she explains 'I don't want to just audit a specific algorithm by itself, I want to audit the algorithm in the context of where it's being used. And compare it to that same context without the algorithm' (Los Angeles Times, 2016). Nicholas Diakopoulos argues that a new accountability perspective is necessary in freedom of information requests. Although there are some examples of successful use of Freedom of Information Act requests to compel disclosure of source codes (Diakopoulos, 2016, p. 59), this is definitely not sufficient to guarantee accountability. He suggests reconsidering FOIA along the lines of Freedom of Information Processing Act which is not so much based on disclosing codes, but allow to submit benchmark datasets the government agency is required to process through its algorithms (Diakopoulos, 2016, p. 59). These are just some examples of efforts to accomplish informational accountability.

Does this imply a profound reorientation of the archival community? Yes and no. No, since it is all about understanding the context. But the efforts to be made to understand the context of creation and use require a reconsideration of the components of the record. The archival community needs to rethink and reconceptualise the essence of a record in a world in which data is ubiquitous, fluid and too abundant to manage and control. If the archival community wants to continue to play a meaningful role in defending informational accountability and

transparency (which includes the historical perspective), a more situated approach is required. Understanding the quality of data and the processing mechanisms of data in situated practices is a prerequisite to be able to play that role. I argue that the apparatus perspective provides a useful framework to understand the archive in situated settings. Only if archivists develop the competences to understand the data assemblages and processing mechanisms in situated practices it will be possible to distinguish between the essential and the residue, the meaningful and the meaningless, the relevant and the extraneous.

Literature

Acker, Amelia (2016). When is a record? A research framework for locating electronic records in infrastructure. In Anne J. Gilliland, Sue McKemmish and Andrew J. Lau (eds.), *Research in the Archival Multiverse* (pp. 288-323). Clayton, VIC: Monash University Publishing. http://dx.doi.org/10.26530/OAPEN_628143

Anderson, Chris (2008). The end of theory: the data deluge makes the scientific method obsolete. *Science*. Retrieved from http://www.wired.com/2008/06/pb-theory/

Batchen, Geoffrey (1998). The art of archiving. In Ingrid Schaffner (ed.), *Deep Storage: Collecting, Storing and Archiving in Art.* Munich and New York: Prestel Verlag and Siemens Kulturprogramm.

Batchen, Geoffrey (2001). *Each Wild Idea. Writing, Photography, History*. Cambridge, MA: MIT Press.

Bearman, David (1989). Archival Methods. In *Archives and Museum Informatics Technical Report #9* (preface). Pittsburgh: Archives and Museum Informatics. Retrieved from http://www.archimuse.com/publishing/archival_methods/#intro

Bowker, Geoffrey C. (2014). The Theory/Data Thing. Commentary. *International Journal of Communication, 8*, 1795-1799.

Cox, Richard J., & David A. Wallace (eds.) (2002). *Archives and the Public Good. Accountability and Records in Modern Society*. Westport-London: Quorum Books.

Boyd, Danah, & Kate Crawford (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662-679.

Canhoto, A.I., & Backhouse, J. (2008). General Description of the Process of Behavioural Profiling. In M. Hildebrandt and S Gutwirth (eds.), *Profiling the European Citizen: Cross-disciplinary perspective* (pp 47-63). Dordrecht: Springer.

Calvino, Italo (1993). La poubelle agréée. In I. Calvino, *The Road to San Giovanni*. New York: Vintage Books.

Cumming, Kate, & Anne Picot (2014). Reinventing appraisal. *Archives and Manuscripts, 42*, 133-145.

Danaher, John (2016). The threat of algocracy: reality, resistance and accommodation. *Philosophy & Technology, 29*, 245-268.

Datta, Anupam, Shyak Sen, & Yair Zick (2016). Algorithmic Transparancy via Quantitive Input Influence: Theory and Experiments with Learning Systems. In *IEEE Symposium on Security and Privacy* (pp. 598-617). Retrieved from http://www.ieee-security.org/TC/SP2016/papers/0824a598.pdf

Deleuze, Gilles, & Claire Parnet (2007). *Dialogues II*. New York: Columbia University Press.

Derrida, Jacques (1996). *Archive Fever. A Freudian Impression*. Chicago / London: University of Chicago Press.

Diakopoulos, Nicholas (2014). *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. New York: Tow Center for Digital Journalism. Retrieved from https://towcenter.org/research/algorithmic-accountability-on-the-investigation-of-black-boxes-2/

Diakopoulos, Nicholas (2016). Accountability in Algorithmic Decision Making. *Communications of the ACM, 59*(2), 56-62.

Diebold, Francis X. (2012). *A personal perspective on the origin(s) and development of "Big Data": The Phenomenon, the Term and the Discipline*. Retrieved from http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf

Floridi, L. (2014). *The 4th Revolution. How the infosphere is reshaping human reality*. Oxford: Oxford University Press.

Floridi, L. (ed.) (2015). *The Onlife Manifesto. Being Human in a Hyperconnected Era*. Heidelberg: Springer.

Gabrys, Jennifer (2011). *Digital Rubbish: A natural history of electronics*. Ann Arbor, MI: University of Michigan Press. http://dx.doi.org/10.3998/dcbooks.9380304.0001.001

Giannachi, Gabriella (2016). *Archive Everything. Mapping the Everyday. Cambridge*, MA: MIT Press.

Hey, Tony, Stewart Tansley, & Kristin Tolle (2009). *The Fourth Paradigm. Data-intensive Scientific Discovery*. Redmond: Microsoft Research. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf

Introna, Lukas D. (2016). Algorithms, governance and governmentality: on governing Academic Writing. *Science, Technology & Human Values, 41*(1) 17-49. https://doi.org/10.1177/0162243915608948

Janssen, Marijn, & George Kuk (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly, 33*, 371-377.

Jimerson, Randall C. (2009). *Archives and Power. Memory, Accountability and Social Justice*. Chicago: Society of American Archivists.

Joh, Elizabeth E. (2015). Policing by numbers: Big Data and the fourth amendment. *Washington Law Review, 89*(35), 35-68.

Ketelaar, Eric (1999). Archivalisation and archiving. *Archives and Manuscripts, 27*, 54-61.

Knobel, Cory Philip (2010). *Ontic Occlusion and Exposure in Sociotechnical Systems*. Dissertation at the University of Michigan. Retrieved from https://oatd.org/oatd/record?recordid=handle%5C%3A2027.42%5C%2F78763

Kroll, Joshua Alexander (2015). *Accountable Algorithms*. Dissertation at the Princeton University. Retrieved from http://dataspace.princeton.edu/jspui/handle/88435/dsp014b29b837r

Latour, Bruno (2009) Tarde's idea of quantification. In Matei Candea (ed.), *The social after Gabriel Tarde: debates and assessments* (pp.187-202). London: Routledge. Retrieved from http://www.bruno-latour.fr/sites/default/files/116-CANDEA-TARDE-FR.pdf

Los Angeles Times (2016), 30 December. Retrieved from http://www.latimes.com/books/jacketcopy/la-ca-jc-cathy-oneil-20161229-story.html

MacNeil, Heather (2007). Archival Theory and Practice: between two paradigms. *Archives & Social Studies: A journal of Interdisciplinary Research, 1*, 517-545.

Manoff, Marlene (2004). Theories of the Archive from Across the Disciplines. *Libraries and the Academy, 4*(1), 9-25.

Mayer-Schönberger, Viktor, & Kenneth Cukier (2013). *Big Data: a revolution that will transform how we live, work and think*. Boston / New York: Houghton Mifflin Harcourt.

Mason, Scott (2016). *Nature of Knowledge*. Delhi: The centre for Internet & Society. Retrieved from http://cis-india.org/internet-governance/blog/nature-of-knowledge#_ftnref19.

McQuillan, Dan (2016). Algorithmic paranoia and the convivial alternative. *Big Data & Society, 3*(2), 1-12.
Morozov, Evgeny (2014). *To save everything, click here. Technology, solutionism and the urge to fix problems that don't exist*. London: Penguin.
Mul, Jos de (2002). *Cyberspace Odyssee*. Kampen: Klement.
O'Neil, Cathy (2016). *Weapons of Math Destruction. How Big Data increases inequality and threatens democracy*. New York: Crown.
Packer, Jeremy (2010). What is an Archive?: An Apparatus Model for Communications and Media History. *The Communication Review, 13*, 88-104.
Rodrigues, Nuno (2016). Algorithmic Governmentality, Smart Cities and Spatial Justice. *Justice Spatiale | Spatial Justice, 10*(July). Retrieved from http://www.jssj.org
Rose, Karen, Scott Eldridge, & Lyman Chapin (2015). *The Internet of Things: an overview. Understanding the issues and challenges of a more connected world*. Geneva (SW) / Reston, VIC: The Internet Society. Retrieved from https://www.internetsociety.org/sites/default/files/ISOC-IoT-Overview-20151014_0.pdf
Røssaak, Eivind (2016). Memory and Media. Archival tasks in the age of algorithms. In *Schetsboek Digitale Onderzoeksomgeving en Dienstverlening. Van vraag naar experiment*. Den Haag: Stichting Archief Publicaties.
Rouvroy, Antoinette, & Thomas Berns (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux, 1*(177), 163-196.
Scott, James C. (1998). *Seeing like a State. How certain schemes to improve the human condition have failed*. New Haven / London: Yale University Press.
Thomassen, Theo (1999). Een korte introductie in de archivistiek. In P.J. Horsman, F.C.J. Ketelaar and T.H.P.M. Thomassen (eds.), *Paradigma. Naar een nieuw paradigm in de archivistiek* (pp. 11-20). Den Haag: Stichting Archief Publicaties). (Translated in English (2000): A first Introduction in Archival Science. Archival Science, 1, 373-385.)
Tokmetzis, Dimitri (2012). *De Digitale Schaduw. Hoe het verlies van privacy en de opkomst van digitale profielen uw leven beïnvloeden*. Houten / Antwerpen: Spectrum.
Upward, Frank, Barbara Reed, Gillian Oliver, & Joanne Evans (2013). Recordkeeping informatics: re-figuring a discipline in crisis with a single-minded approach. *Records Management Journal, 23*(1), 27-50.
Villasenor, John (2011). *Recording Everything: Digital Storage as an Enabler of Authoritarian Governments*. (Center for Technology Innovation (Brookings)). Retrieved from https://www.brookings.edu/research/recording-everything-digital-storage-as-an-enabler-of-authoritarian-governments/
Wetenschappelijke Raad voor het Regeringsbeleid [Scientific Council for Government Policy] (2011). *iGovernment*. Amsterdam: Amsterdam University Press.
Wetenschappelijke Raad voor het Regeringsbeleid [Scientific Council for Government Policy] (2016). *Big Data in een vrije en veilige samenleving*. Den Haag: WRR. Retrieved from https://www.wrr.nl/publicaties/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving
Yeo, Geoffrey (2007). Concepts of Record (1): evidence, information, and persistent representations. *The American Archivist, 70*, 315-345.