"The data dilemma: who is in control?"

Tove Engvall, Mid Sweden University

tove.engvall@miun.se

Introduction

Digitalization enables business and investment activities to be carried out online on a global scale. This brings possibilities, but also challenges regular processes to ensure accountability between stakeholders, as well as citizens' rights. The online environment makes fraudulent activities easy, and actors vulnerable; it is difficult to know who can be trusted (Engvall, 2016). The *Global economic crime survey* 2016 (PwC, 2016) indicates new paths for economic crimes, a complex threat landscape in a fast-paced global marketplace. Cybercrime is currently the second most common economic crime, and fraud a common problem. The costs of economic crimes do not only include monetary loss, but also investigations, damage of reputation and morale, and impacts on long-term business performance. Digital technologies enable faster connection and on a wider scale than before, which is both a possibility and a risk. The digital environment is complex and the development fast. Democratic institutions have to keep up with the development in order to maintain democratic functions in society. Big Data is discussed to provide means for meeting some of these challenges.

In a fast-changing world, it is important to have mechanisms for creating trust within societies. This is what records and archives management have been doing for ages, even though the context and technologies have changed over time. A key concern is to ensure trustworthy, authentic and reliable evidential records of agents' activities (ISO 15489-1:2016). Records are created in course of activities, and provide evidence of these activities (Yeo, 2011). Records can be used to hold actors accountable and provide evidence of criminal or unfair behaviour and is crucial to maintain justice and rule of law. As the use of data grows, so does the need to ensure its trustworthiness and preservation, which means that there is need for new approaches to apply in records management (Coleman, Lemieux, Stone & Yeo, 2011). Records can provide evidence in business disputes and improve litigation readiness (Ardern, 2011). Good management of evidences of business operations can also strengthen consumers' rights and promote honest operations on the market in general. The idea is also that more data of what takes place in the financial market can improve means for monitoring, foresee risks, identify breaks of regulation, and increase knowledge that can be used to mitigate a new financial crisis and promote financial stability. In today's online environment, a great part of records of transactions tend to be in the form of data.

Research on the financial crisis 2007-2008 indicates that some of the reasons were connected to a lack of control, such as liberalization of government regulation, governance issues within financial firms and insufficient records management, along with patterns on the market. A robust records creation and recordkeeping system is crucial for operations and ability to manage risks on the financial market, as well as internal operation of a business (Coleman, Lemieux, Stone & Yeo, 2011). Records support regulatory functions in financial institutions and are often central in compliance work. There are different levels of regulations of the financial markets, and there is a move towards EU harmonization, with MiFID (The Markets in Financial Instruments Directive) as a keystone of EU financial law, which also includes requirements on recordkeeping (Herbst & Lovegrove, 2011). As will be discussed in the interviews that has been conducted for this paper, new regulations are coming into force in 2018, including new requirements to provide records (in form of data). Even though there are new means for creating, gathering, analysing and sharing information, there are also challenges with faster and

more complex and expensive technological tools, which again raises challenges regarding control of the information. It seems as if any improvement brings new challenges. It seems as if gathering, analysing and exchanging data can be used to solve some of the problems we face in the global market environment, and are closely connected to increasing control. However, this also involves risks and ethical concerns that have to be addressed. Large quantities of data also means a concentration of power, and the question is who is in control of the data?

Objectives and method

As digitalization in the domain of the financial markets evolves, and activities are carried out online, regular processes for democratic control, accountability and ensuring citizens' and different actors' rights are challenged and at risk. In this context, what are the possibilities of using big data to keep up with the technological development and still ensure these qualities? Moreover, what are the risks?

Semistructured interviews have been made with employees at public authorities, at EU level and national level. Respondents include the Head of the ESRB Secretariat at the European Systemic Risk Board, employees at the national financial supervisory authorities in three different EU countries (one from each authority), and two employees at a national company registration office. Further interviews with other authorities could be of value, for example tax agencies and economic crime offices, as well as political representatives. However, this could be a continuation in future studies. Interviews were carried out either by telephone, e-mail or in person. There were also authorities that were unable to participate in the study. Initial contact was made by e-mail, followed by phone calls, with the exception of one case where the reply was sent by e-mail. Research articles were primarily searched for in the IEEE and Google Scholar databases, and the Records Management Journal. Search terms that was used were for example *big data* and *financial market*, *machine learning*, as well as *Computational Archival Science*.

Related research

This section aims to address different aspects related to big data and means for control of financial activities, as well as the connection to archives and information science. It includes articles about big data, means for processing and analysing big data, Computational Archival Science (CAS), eDiscovery and Digital Records Forensic. Because of time and space limitations, this will not be an in-depth investigation into these fields, but a taste of what it could mean, which can be explored further in future research.

Big data, data mining, machine learning and Visual Analytics

Big data can be explained as data of big volume, variety and velocity (speed of in and out data), which requires more than the commonly used tools to capture, manage and analyse the data (Lemieux, Gormly, Rowledge, 2014). "*Data mining* is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns" (Witten, Frank, Hall & Pal, 2017, p. xxiii). Data mining is about discovering patterns in data in an automated way, that also provide some value. For example, it is used for environmental purposes, medication, consumer choices, cyber security aspects, banks and credit assessments, financial market monitoring and more (Witten, Frank, Hall & Pal, 2017). Big data has also been used in democratic rule-making processes, to increase transparency and innovate consultation with citizens (Lemieux, 2016). *Machine learning* is about computers' ability to answer questions, and tends to be directed towards prediction of decision-making. Learning in this context rather refers to performance and practical learning rather than theoretical knowledge. The idea is to explain patterns and make them understandable, so that they can serve as a basis for

prediction (Witten, Frank, Hall & Pal, 2017). Humphries (2017) argues that machine learning is a kind of AI, and can be explained as "the capacity of computers to learn without being explicitly programmed. Machine learning involves computers taking data and algorithms as inputs to develop models of algorithms that apply this learning to novel data" (Humphries, 2017).[1] There are practical tools that can be used to extract useful information from raw data, but it is also important to recognize that data is imperfect; it can be incomplete and not completely reliable (Witten, Frank, Hall & Pal, 2017). There are often challenges relating to the quality of the data, as well as making sense of it (Lemieux, Gormly, Rowledge, 2014).

It is also important to consider ethical implications. In the online environment, everything people do is recorded. Patterns in behaviours can be used for commercial, research or political purposes, and there is a big commercial hype around machine learning. What is new about this technology is the increased possibilities for discovering patterns and analysis. This has to be treated with responsibility. Discrimination due to, for example, ethnicity or socio-economic status is one risk, and it is important to consider how data can be used, and what people have the right to know of how the data they provide in different contexts are managed and what it will be used for (Witten, Frank, Hall & Pal, 2017). One tool for the management, interpretation, and understanding of big volumes of data, is Visual Analytics (VA). VA combines computational capabilities with graphical representations; it uses large volumes of data and enables interactive analysis. At the centre of concern, however, is good management of records in place (Lemieux & Baron, 2011).

In the financial domain, machine learning can for example be applied to prediction of financial crisis, and are used at banks for prediction of bankruptcy and credit scoring. It is about predicting a loan applicant's level of risk for a bank (Lin, Hu & Tsai, 2012). Machine learning techniques can also be used in different applications, for risk management, trading and portfolio selection. Efficient analysis of data is important to understand systemic risk and to develop frameworks that can include data from different sources and of different characteristics. Regulatory compliance, fast development of new complex financial innovations and risk management requirements has to be considered (Serguieva, 2014). Financial markets can be seen as service systems that creates value for stakeholders, but also includes risks in which technology plays an important part. Algorithms in automated trading can have an unexpected behaviour, resulting in shock and domino effects. There are also inequalities in trading due to access to different technologies and volumes to trade with. There are four aspects that are important to consider in a monitoring and surveillance system according to the authors: to address not just the parts, but also the holistic perspective; recognize ongoing changes and development of new tools and techniques; agents acting in unpredictable ways; and collaboration and coordination between actors that need to be included in order to detect market manipulation. For these reasons, a combination of behavioural analysis and economic analysis is suggested (Diaz, Theodoulidis & Abioye, 2013).

Fraud detection is another field where data mining can be used. For example, detection of anomalies can be indicated to prevent credit card or identity theft, or market manipulation. Being able to analyse time series, including for long periods of time, and be able to include high frequency trading with thousands of transactions recorded per second is an advantage (Golmohammadi & Zaiane, 2015). Golmohammadi, Zaiane and Díaz (2014) argue that better tools are needed to detect fraud, suspicious transactions and market manipulation, and machine learning is one way to improve these tools. Large volumes of money are lost because of fraud, which is a big cost for society. Data mining and learning algorithms could be used to detect market manipulation, but the management and heterogeneity of the data involve certain challenges (Golmohammadi, Zaiane & Díaz, 2014). According to Huang, Liang, and Nguyen (2009), further measures need to be taken to ensure security, prevent fraud and

[1] https://futureproof.records.nsw.gov.au/machine-learning-and-records-management/

attacks; they suggest analysis of trading networks with behaviour-driven visual analysis. They argue that methods relying on AI often produce false alarms, and that automated monitoring systems are not enough. They also highlight the advantage of including historical data in analysis, to make historical comparisons as well. They claim that a smart surveillance system of the market is crucial in order to ensure fair trading, but it will require high performance technology. Another area of fraud where big data can be applied is tax evasion (Abrantes, Ferraz & Abrantes, 2016). Governments suffer economic losses due to tax evasion. In the UK, the estimated tax losses between 2012 and 2013 was £ 15.4 billion, and the Brazilian government estimate a loss of R$ 415.1 billion in 2015. According to the authors, tax evasion is directly linked to the probability of detection, and the use of big data can be a way to improve its efficiency. It can be used to increase the volume of audits, improve tax evasion detection, complex fraud investigations and means for analysis of company and individual data. However, it also requires innovative information processing that enables management of big volumes, variety of data as well as velocity in analysis (Abrantes, Ferraz & Abrantes, 2016). PwC (2016) also raise other concerns to address fraud, such as the organizational culture and governance. It is also a question of what kind of society we want, and to also work with aspects of trust and develop a balance between trust and control.

Computational Archival Science

Because of the challenges connected to the management of big volumes of data, in both archival and computational fields, researchers are looking for new ways to make data useful and to ensure its quality and trustworthiness. Some of the challenges identified concern long-term preservation, interpretation, trustworthiness and means for analysis. The development of a trans-discipline called Computational Archival Science (CAS) has been suggested; it would apply both efficient computer processing techniques for data management and archival methods to ensure authenticity and reliability as well as long-term preservation of data. It would include development of effective capturing, management and use of data, along with considerations of ethical, security and privacy issues, and in addition address societal and organizational concerns for trustworthy and authentic records. Both computational and archival science and concepts would be used to develop an integrated theoretical foundation (Marciano, Lemieux, Hedges, Esteva, Underwood, Kurtz & Conrad, 2017). Areas of concern would for example be digital curation and cyber infrastructure. Digital curation includes activities to maintain digital materials over time, such as activities concerning long-term preservation of data, and adding value to the data in order to facilitate interpretation and exploration of the material. New emerging technologies, with new needs for preservation of information are another field of concern (Marciano, Lemieux, Hedges, Esteva, Underwood, Kurtz & Conrad, 2017). Due to regulations, not having sufficient data available could involve a high cost, and the value of information might even increase over time, with possibilities for different kinds of analysis (Viana & Sato, 2014). Data analytical techniques can be used for archival description and to facilitate access to records and archives. Initiatives of distributed, linked and interconnected archival collections have been tried in the UK, with an aim to facilitate research and exploration of the material. It requires new ways of thinking and collaboration between archival and computer scientists. New ways for users to access archival material could be developed, yet it might be even more crucial to maintain the provenance in a complex environment (Ranade, 2016). With publication of data openly accessible online, there is an increased need to control data creation and management, as well as documentation of data, in order to ensure its trustworthiness (Lemieux, Gormly & Rowledge, 2014). A challenge for many archivists and records managers is the mindset around digital material. Bunn (2016) argues that archivists need to think differently and go beyond the "paper mind". Computational methods and information processing capabilities are not fully used because of the way people think. Correspondence between the external reality and "what goes on within the human head" ought to be addressed (Bunn, 2016, p. 3243). At the State Archives and Records NSW, machine learning technologies are tested to automate records classification and disposal. Machine learning has also been tested to assist in appraisal and sensitivity review processes, using eDiscovery tools. It can

be useful in digital transfers, but human action is still required in the process. There are commercial machine learning products on the market, used to for example classify data and identify patterns. The risks have to be addressed, and it has to be proven to work, in order to be trusted (Humphries, 2017). Lemieux (2012) discusses the possibilities of using Visual Analytics (VA) in the archival domain, for example to facilitate arrangement and description, manage unstructured records, and analyse archival material (Lemieux, 2015). It is important to ensure both the trustworthiness and transparency of the data, as well as the transparency in the process of analysis and development of tools and methods, to facilitate for the user to assess the results. This includes preservation of data, ensuring qualities of data, work with metadata, policies, procedures and responsibilities (Lemieux, Gormly & Rowledge, 2014). In addition, Maemura, Becker and Milligan (2014) address the importance of transparency in automated processing of data. They discuss it in a context of research in web archives, and a need to develop and explain methodological frameworks adapted to automated processes. This in order to address questions of provenance, and for users to be able to assess the adequacy of the findings.

EDiscovery, Digital Records Forensics and use of data to mitigate economic crime

Another field related to control of data is eDiscovery, both as a way to increase readiness for risk of crime, as well as possibilities of crime investigation in the digital environment. "Electronic discovery, or eDiscovery (…) is a process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a legal case" (Lawton, Stacey & Dodd, 2014). According to a report from PwC (2016), a company working with eDiscovery, economic crime is changing and detection and control mechanisms are not keeping up with this change. They argue that "the burden of preventing, protecting and responding to economic crime rests firmly with organisations themselves" (PwC, 2016, p. 8). This also applies to individuals who act in the online context in for example online trade (Engvall, 2016). Cybercrime is increasing, and is reported as the second most reported economic crime in 2016, in addition to which fraud is also common. Fast interconnections increase vulnerabilities and can have impact on a wider scale. However, few actors have a cyber-incident response plan and are not adequately prepared for incidents, and not many have performed a fraud risk assessment (PwC, 2016). They promote the use of big data analytics, but emphasize the need to work with overall strategies: value-based business culture, aligning ethics with decision-making, risk management and governance in the organizations. They also discuss money laundering as a big problem that fund activities like terrorism, corruption, tax evasion, and drug and human trafficking. Authorities detect few of these illicit financial flows, and it is an area in need of improvement (PwC, 2016). In a report about the use of big data in the police in UK (Babuta, 2017), it is stated that big data has been applied for limited tasks, but could be very useful in preventing crimes, foreseeing risks and protecting victims. However, there are also societal concerns regarding surveillance and risks of discrimination. If for example minorities have been disproportionally targeted by police actions in the past, algorithms will disproportionally assess those individuals as risks, with racial discrimination as a possible consequence. More advanced technologies and data analysis are also used by criminals, and in an online environment that easily enables anonymity, measures have to be taken to prevent certain data and technologies to get into the wrong hands (Babuta, 2017). It would be interesting to take a closer look at the United Kingdom. Western Europe has the second most reported economic crimes in the world, with the United Kingdom having the second most reported economic crimes (PwC, 2016, p. 9-10). Conditions for the localization of online economic crimes could be interesting to look further into. Many scam companies in online trade claim to be based in London for example, as well as it has been suggested that the control of business registrations is low in the UK (Engvall, 2016). In a global context, no country is an island; each country affects actors in other countries as well as actors in their own country.

As records provide evidence of activities, they are crucial in processes that search for and manages digital evidence. Duranti (2009) has proposed a development of a body of knowledge called Digital Records Forensics, which combines archival knowledge and means for ensuring authenticity, with

digital forensics methods and concepts (which are also related to eDiscovery). The archival profession provides means for ensuring authenticity and preservation of records, and functions as a neutral third party, or 'trusted custodian'. In the InterPARES project, a trusted custodian has been defined as "a neutral third party who must demonstrate that it has no reason to alter or to allow others to alter the records in its care, and that it has the knowledge required for attesting to, and ensuring the continuing authenticity of, the records." (Duranti, 2009, p. 41). Digital forensics is often used in investigations of digital material, to find evidence of criminal activities. Digital Records Forensics should include design of digital systems that create and maintain trustworthy digital records as evidence, serving accountability purposes over time, and providing means for verification of authenticity in cases of weak assumptions (Duranti, 2009).

Interviews

As mentioned, employees at a number of different public authorities have been interviewed, at European and national level. Including the European Systemic Risk Board, financial supervisory authorities in three European countries, as well as one European national company registration office.

The European Systemic Risk Board (ESRB)

The European Systemic Risk Board (ESRB) aims to "oversee the financial system of the European Union (EU) and prevent and mitigate systemic risk" (ESRB, 2017). It was established 2010 as a response to the financial crisis to increase the supervision of the financial system, to rebuild trust in the financial system, and strengthen the protection of European citizens (ESRB, 2017). The ESRB works in close cooperation with the European Securities and Markets Authority (ESMA), which works for the protection of investors and to promote stability of the European Union financial markets. The ESMA provides registers and statistical financial data to regulators, market actors and the public (in the framework of EU legislation), as well as warnings and information to investors (ESMA, 2017). The ESRB and the ESMA have introduced big data in their work. The ESRB, for instance, works with two large datasets: EMIR and AIFMD. EMIR is an EU directive that requires financial firms to report data about transactions in derivative markets. It includes 50 million datasets per day from transactions on the derivative market. The transaction data can be analysed to identify economic issues and problems that require discussion, while maintaining confidentiality according to EU legislation. AIFMD (the Alternative Investment Fund Manager Directive) is a directive that aims to regulate the most speculative investment funds. An objective of the directive is to improve access to the financial market operations for public regulators, following detailed technical standards of what companies need to provide. Ensuring data quality, completeness of data, and correction of bad reporting, i.e. reporting that does not make sense, are some of the issues they have to pay attention to, and which statisticians are working on. To be able to analyse issues, it is important to have long series, and these series have to be stored and preserved. In fact, identification of dependencies requires as many observations as possible, to enable, for example, cross-data analysis. The collection of data under EMIR and AIFMD is a requirement that has been agreed upon internationally in the G20 group. Before this legislation entered into force, some of the financial markets mentioned above were like a black box. As stated by one respondent, this will no longer be the case. However, it is too soon to say anything about the result of the analysis, since tests are still being carried out. As for the capacity to collect and store such a huge dataset, statisticians and archivists working at the ECB (European Central Bank) are supporting the ESRB, and the technology to this end is being established. The ESRB looks at whether systemic risks are materializing in financial markets; in analysis of data they use, for instance, indications such as concentrations and locations of the market, violent price movements and structural problems. As financial bubbles can emerge on the market, they sometimes need to be balanced by macroprudential policy. For example, collective behaviour can create instabilities. The ESRB watches for market developments that need to be discussed by macroprudential supervisors, and may give warnings and recommendations concerning market stability. The ESMA (European Securities and

Markets Authority) share the same data, but use it for other purposes, such as to avoid market abuse and market manipulation. They also collaborate and provide access to the data for other institutions, such as National Financial Authorities. (R1)

National Financial Supervisory Authority, Country A

The National Financial Supervisory Authority (R2) "uses data analysis methods to detect possible violations of regulations. For example, (the Authority) routinely analyses trading data, and—when necessary—order data on securities transactions which credit and financial services institutions have to report for the purposes of monitoring compliance with the prohibition of market abuse. In this context, (the Authority) also monitors all ad hoc notifications of listed companies to detect possible breaches. For the analysis, (the Authority) uses manual, automated and semi-automated big data and data mining solutions and specialized software. This software can be proprietary in-house software as well as commercial software" (R2, 2017). This is used to analyse transaction data, and it contains data from different sources. They continuously review and improve methods and tools. Challenges that they face relate to, for example, system performance, developments on the market such as high frequency and algorithmic trading, and more complex information management. As market abuse schemes become more complex and advanced, they also become more technically complicated to map and incorporate into the surveillance system and tools. As mentioned, they continuously improve and apply new tools, and machine learning could be used in surveillance solutions to predict trends in market abuse. The law regarding market abuse is primarily repressive, but it also has preventive dimensions such as reporting of manager's transactions to prevent insider trading. The data is considered relevant for 5 years, due to regulation periods. The requirements in MiFIR about transaction reporting from market participants will be integrated in their systems. It requires more detailed transaction data than MiFID I, which is why it will provide increased opportunities for data analysis. (R2) From the Division for Reporting Requirements, they have specialists to ensure data quality, data plausibility and data integrity. There are also database designers who are responsible for indexing and ordering data, information integration and preservation possibilities as well as performance issues and replication technologies. They also have a special data protection officer, responsible for sensitivity concerns. (R2)

National Financial Supervisory Authority, Country B

The national financial supervisory authority (R3) uses a lot of data to survey all financial markets, to detect and prevent market abuse, manipulation and crime, and to enforce policy for market violation and market abuse. They observe the market to make sure it is working properly. They carry out real-time monitoring (which is not done in all European countries), as well as non-real time monitoring. They carry out the monitoring themselves, but also have partnerships with companies that provide certain data management services. They make big investments in ensuring that data is accurate, and they are quite satisfied with the data they gather. There are some regulations about data provision, and new EU regulations (MiFID2 and MiFIR) will enter into force in January 2018. MiFIR includes detailed requirements of records that financial firms need to provide about transactions, and several companies are making IT investments in order to ensure compliance with regulations, and to be able to provide data in correct format. Regulations impose investment firms to store data for at least 5 years; the authority keeps its own database for an even longer period of time, up to 15 years. This is because they believe it can be used to make more long-term analysis of patterns, movements and more. There are also many enquiries on the data, for example from other authorities. The conditions for archiving data is challenging, because of the big volume of data. The demands for data has exploded over the last 10 years, and current and future market developments are crucial. Therefore, it is important to find a reliable and proper solution for long-term preservation of data. There are also several old databases, and it is a challenge to maintain the quality and security of the information. IT-specialists deal with these questions, and archivists appear to not be involved.

They use several tools to interpret the data, focusing on two main tools: data mining, which can be used to make different types of analysis, for example, to understand market movements, price movements and other patterns. The other tool provides notifications of suspicious patterns, which can then be assessed as to whether it should be further investigated. MiFID2/MIFIR will make it easier to make investigations, since it will provide data about all transactions on the market in their own database, i.e. there will be no need to request the information from the companies. There are also collaborations between national authorities and interchanges of data. Collection of data will include companies that operate within the regulation; those who are not regulated are not included (R3). This means that clients using unregulated companies are unprotected, and their transactions not supervised. As to what the respondent (R3) sees as major challenges ahead, data management is the primary concern. Big volumes need to be managed every day, indexed properly and captured in a well-built database ensuring quality and ability to respond to enquiries in an efficient way, and to be a reliable system. (R3)

National Financial Supervisory Authority, Country C
The respondent at this authority works with the banking sector, which is why it is a focus of concern in this interview. The Authority receives very detailed information from banks regarding capital and liquidity situations. The information concerns, for example, annual reporting, balance sheets, income statements and credit portfolios. The data is like a blood test, and the Authority uses indicators for analysis. They look at things like market risk, credit risk, liquidity risk, return, and costs in relation to incomes. The purpose is to identify activities that can be a risk, and if further investigation is needed. In addition, they act in accordance with regulations, for example making sure that the capital base is high enough, and ensuring the quality of the capital. The analysis of reported data related to the indicators are automated, and they use around 200 different indicators. After the analysis of the indicators, an assessment will be made as to whether further analysis is necessary. They also look at reports, and have personal contact with the banks. Reporting from the banks is done in a specific interface. The banks logs in and goes through a number of validation steps. The authority go through a range of controls when the data has been submitted. There is also one step where the data is used, and where anomalies are detected. For example, if they note fast changes in capital. It is possible something has happened at the bank, or they could have entered the wrong currency in their report. They also submit some of the information to the European Bank Authority (EBA), who carry out further controls. If someone wants to manipulate the data, every number reported has to be changed, which would make it a full-scale fraud. Sharing information with other authorities could be valuable, for example with the police and the National Company Registration Office. It would provide them with indications of suspicious activities, such as money laundering and economic crimes. The assessment of applications from new banks would also be more efficient if information was more easily accessible.

The digitalization brings with it new challenges. While the same services are carried out, it is in a new, digital context. However, it can sometimes change a business model. New initiatives like crowdfunding and block chain technology challenge the banks' position in different fields. The authority observes trends, and if negative consequences are noted, regulation measures should be considered. New technologies means new possibilities, but also challenges. If there is a power failure, if you lose your mobile phone or your computer shuts down, which could and does happen, then additional issues of security have to be considered.

National Company Registration Office
The Authority registers companies and receives annual reports. There are company documents in their archive, which according to law have to be preserved. It primarily includes basic information about companies, as well as annual reports. They are currently considering how to provide annual reports electronically, and looking at the format ixbrl, which is soon going to be a standard. By using a

standard format, it becomes easier to exchange information with different stakeholders. They argue that using digital reporting will provide them with better control, and it will enable big data analysis. Combining basic data about companies with data in annual reports, further analysis will be possible. They are also looking at possibilities to retrieve information about real principal. The EU requests information about actors that control and have influence of companies. Analysis of these parts, of basic data, annual reports and real principal, would make it more difficult for criminals. Digital management also means a need for better software, which would facilitate reporting. Using electronic receipts and accounting, they could increase the speed and transparency of the accounting reporting. They are also considering possibilities to replace the annual reporting with more frequent reporting, in which case it would be more transparent and they could be more responsive. In the Netherlands, they use a standardized format for business reporting. An advantage of this is that companies frequently improve their interest rates at banks, who can make more precise risk assessments by using big data analysis. The respondent believed that when routines for electronic reporting have been introduced, including a standardized format, and accounting software has been improved, this could have some advantages. The banks would be able to make better risk assessments, regarding both credit scoring and risks of bankruptcy. There are, however, risks of increased discrimination, where economically weak groups are scored as a higher risk, receiving a higher interest rate. In personal customer relations, other aspects can be considered as well, and trust is a crucial factor.

Also other authorities require reporting. Today the Tax Agency in the country have their own format, but if they would use the same as to the National Company Registration Office, it would make it easier to exchange information, and also increase transparency, and provide means for further analysis. With improved means for analysis, both at banks and authorities, suspicious activities and risks could be detected at an earlier stage, and it could be used in research. Above all, it is an advantage for the public, as it implies increased transparency. Regarding openness of the information, there are also risks. Analysis can be made for criminal purposes as well, which could put companies in a vulnerable position. This has to be carefully considered. Certain information should only be shared with the police for example, and not be accessible to the public in general. (R5) They have a group looking at ways to control reports from companies, to increase possibilities to prevent crime. The idea is to have means to identify patterns that indicate that something is suspicious according to different parameters, which in turn could indicate a crime. They are in the process of investigation, and have no results as of yet. (R6)

Discussion

Digitalization means that activities are carried out in ways that challenge traditional processes and mechanisms for accountability, democratic and institutional control as well as trust. The use of big data analysis is a way to increase control in different areas, such as the monitoring of the financial markets, indication of fraud, assessment of credit risk, and to facilitate crime prevention and investigation, and more. Collaboration between public authorities, as well as between authorities and market actors, can enable a more efficient information management that can promote a fairer, more transparent and accountable financial market. In addition, it can contribute to strengthen democratic functions, and build a basis of trust and common values, in which trustworthy records will be central. Due to the challenges of volume, velocity and variety of data that needs to be managed, Computational Archival Science, eDiscovery and Digital Records Forensics could be useful for further exploration. Still, the technological development is very fast, and democracy takes time, and different approaches to rulemaking should be considered, both to make regulation more responsive, as well as legitimate. As Lemieux (2016) discussed, increased participation can improve its quality and legitimacy, and is important that it is performed in a transparent manner.

As new technologies emerge, so do new challenges that need to be managed. Having control of records is becoming more complex, as well as more important, due to needs for accountability and because money is actually represented as records. Control of records management will also increase control of money. MiFID2 and MiFIR regulations is a step forward, and it can increase the transparency of the financial market. However, there are also many actors on the market who are not regulated, and so will not be included in this initiative. This is probably also where most criminal activities occur. As stated by a trading business leader (in a conversation in the study of online trade), those who follow the law have a lot of work following the rules, while the criminals go free. Part of the digital reality we are living in is the continuous technological development, where new initiatives occur that try to go around systems and regulations, and this is something we need to deal with. For this reason, authentic and reliable records and records and archives managements as a basis in a trustworthy infrastructure can provide a more resilient environment, and even though technologies change, the information maintain certain qualities. With an increasingly changing environment, the need for reliability will be crucial. As different users make more choices, and with an increasing responsibility for their choices, means for assessing trustworthiness is something to consider. This in order for people to more easily make assessments of who they want to collaborate with, for example, and what their goals are, to avoid funding terrorism, trafficking or other criminal activities without knowing.

Ethical considerations

The use of big data for 'good' also brings with it risks and ethical issues that have to be considered, such as issues of surveillance, concentration of power, control and privacy. The risk of biased data, and technologically instituted discrimination and structural abuse, have to be considered. There is also a risk of an over-reliance on technology and data; we have to remember that nothing is perfect, data can be inaccurate and we have to maintain empathy, use our consciousness and look beyond technological structures when required. Technology has to be developed to support human goals, and the values that are built into the systems are crucial to consider. It is still humans that specify the task, define data input and output, develop the algorithms and more, and these processes need to be transparent and accountable (Humphries, 2017). The amount of data about people and activities managed online is a liability that can be used for large-scale crimes, such as fraud. This is not a risk for individuals and businesses only, but also entire societies. The ruination of a large number of citizens in a country can generate huge economic damage in a country's economy. Security services that can be used to ensure that personal data does not end up on the black market (MySafety for example), greater readiness for cyber risks, as well as psychological defence and increasing public awareness of risks, are activities that can be considered. So far, the consequences of gathering all data of human activities are unknown, as is its social implications, impact on individuals, and trust between people (Lambiotte & Kosinski, 2014). What control it gives to those in charge of the information is also known. What are the means to influence the behaviour of people and propaganda using today's technologies? What underlying values are promoted, and in whose interests? What happens if people with racist values are elected to political positions? We might need some kind of emergency plan, to ensure that it is not used for oppression. As has been mentioned, data can be used both for good and for bad. It can further improve democracy and citizens' rights, transparency and accountability, and prevent and detect crimes—as well as shape new forms of control, power, and risks of abuse and crime, and control of people.

References

Abrantes, P. C., Ferraz, F. & Abrantes, P. C. (2016). Big Data Applied to Tax Evasion Detection. *2016 International Conference on Computational Science and Computational Intelligence.*

Ardern, C. (2011). Discovery and records management. In: Coleman, L., Lemieux, V. L., Stone, R. & Yeo, G. "*Managing Records in Global Financial Markets ensuring compliance and mitigating risk"*, (pp. 165-178). Facet publishing, London, UK

Avlan H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal. (2017). *Data mining. Practical Machine Learning Tools and Techniques*. Elsevier, Cambridge, USA

Babuta, A. (2017). Big Data and Policing. An Assessment of Law Enforcement Requirements, Expectations and Priorities. *Royal United Services Institute for Defence and Security Studies (RUSI),* United Kingdom

Bunn, J. (2016). Mind the Explanatory Gap: Quality from Quantity. *2016 IEEE International Conference on Big Data (Big Data)*

Coleman, L., Lemieux, V. L., Stone, R. & Yeo, G. "*Managing Records in Global Financial Markets ensuring compliance and mitigating risk"*, Facet publishing, London, UK

Diaz, D., Theodoulidis, B. & Abioye, O. E. (2013). Monitoring and Surveillance Systems for Financial Markets. A Service System Perspective. *IEEE International Conference on Business Informatics.*

Duranti, L. (2009). From Digital Diplomatics to Digital Records Forensics. *Archivaria, 68 (Fall 2009),* pp. 39-66

Engvall, T. (2016). Fear, Greed and Lack of Trust in Online Financial Trade. *Journal of Administrative Sciences and Technology. Vol. 2017*, Article ID 106163, 10 pages.
DOI: 10.5171/2017.106163

European Systemic Risk Board (ESRB). (2017). https://www.esrb.europa.eu/about/html/index.en.html (Accessed 2017-11-01)

European Securities and Markets Authority (ESMA). (2017). https://www.esma.europa.eu/ (Accessed 2017-11-01)

Golmohammadi, K. & Zaiane, O. R. (2015). Time Series Contextual Anomaly Detection for Detecting Market Manipulation in Stock Market. *International Conference on Data Science and Advanced Analytics (DSAA)*

Golmohammadi, K., Zaiane, O. R. & Díaz, D. (2014). Detecting Stock Market Manipulation using Supervised Learning Algorithms. *International Conference on Data Science and Advanced Analytics (DSAA)*

Grant, S., Marciano, R., Ndiaye, P., Shawgo, K. E. & Heard, J. (2013). The Human Face of Crowdsourcing: A Citizen-led Crowdsourcing Case Study, *IEEE International Conference on Big Data*

Herbst, J. & Lovegrove, S. (2011). Moves towards a common regulatory framework for financial services in the European Union. In: Coleman, L., Lemieux, V. L., Stone, R. & Yeo, G. "*Managing Records in Global Financial Markets ensuring compliance and mitigating risk"*, (pp. 41-59). Facet publishing, London, UK

Huang, M. L., Liang, J. & Nguyen, Q. V. (2009). A Visualization Approach for Frauds Detection in Financial Market. *13th International Conference Information Visualisation*

Humphries, G. (2017) Machine learning and records management. In: Future proof - protecting our digital future. A State Archives and Records Initiative for the NSW Government. https://futureproof.records.nsw.gov.au/machine-learning-and-records-management/ (Accessed 5 November, 2017).

International Organization for Standardization (2016). ISO 15489-1:2016. Information and documentation – Records management – Part 1: Concepts and principles. *Swedish Standards institute,* Stockholm, Sweden

Lambiotte, R. & Kosinski, M. (2014). Tracking the Digital Footprints of Personality. *Proceedings of the IEEE. Vol. 102, No. 12, December 2014*, pp. 1934-1939. DOI: 10.1109/JPROC.2014.2359054

Lawton, D., Stacey, R. & Dodd, G. (2014). eDiscovery in digital forensic investigations. The Centre for Applied Science and Technology (CAST), Publication Number 32/14. Accessed at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/394779/ediscovery-digital-forensic-investigations-3214.pdf 2017-11-06

Lemieux, V. & Baron, J. R. (2011). Overcoming the digital tsunami in e-discovery: is visual analysis the answer? *Canadian Journal of Law and Technology, Vol. 9, No 1 & 2 (2011)*

Lemieux, V. L. (2012). Using information visualization and visual analytics to achieve a more sustainable future for archives: A survey and critical analysis of some developments. *Comma. Vol. 2012, issue 2*, DOI: 10.3828/comma.2012.2.6

Lemieux, V. L., Gormly, B., Rowledge, L. (2014) "Meeting Big Data challenges with visual analytics: The role of records management", *Records Management Journal, Vol. 24*, Issue: 2, pp.122-141, https://doi.org/10.1108/RMJ-01-2014-0009

Lemieux, V. L. (2015). Visual analytics, cognition and archival arrangement and description: tudying archivists' cognitive tasks to leverage visual thinking for a sustainable archival future. *Archival Science, Vol. 15*, pp. 25-49. DOI 10.1007/s10502-013-9212-y

Lemieux, V. L., (2016). Innovating Good Regulatory Practice using Mixed-Initiative Social Media Analytics and Visualization. *IEEE, International Conference for E-Democracy and Open Government.* DOI 10.1109/CeDEM.2016.38

Maemura, E., Becker, C. & Milligan, I. (2016). Understanding Computational Web Archives Research Methods Using Research Objects. *2016 IEEE International Conference on Big Data (Big Data)*

Marciano, R., Lemieux, V. L., Hedges, M., Esteva, M., Underwood, W., Kurtz, M. & Conrad, M. (2017). "Archival Records and Training in the Age of Big Data." In Jaeger, Paul and Sarin, Lindsay. *Re-Envisioning the MLS* (Emerald Group Publishing, forthcoming 2017).

Pwc (2016). Global Economic Crime Survey 2016. Adjusting the Lens on Economic Crime. Preparation brings opportunity back into focus. Pwc, Stockholm, Sweden

Ranade, S. (2016). Traces through Time. A probabilistic approach to connected archival data. *2016 IEEE International Conference on Big Data (Big Data)*

Serguieva, A. (2014). Systemic Risk Identification, Modelling, Analysis, and Monitoring: An Integrated Approach. *IEEE CIFEr General Chair, and Systemic Risk Panelist. (Financial Computing and Analytics Group. Department of Computer Science, University College London)*

Viana, P. & Sato, L. (2014). A proposal for a Reference Architecture for long-term archiving, preservation and retrieval of Big Data. *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*

Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai (2012). Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews, Vol. 42*, No. 4, July 2012,

Yeo, G. (2011). Introduction to the series. In: Coleman, L., Lemieux, V. L., Stone, R. & Yeo, G. "*Managing Records in Global Financial Markets ensuring compliance and mitigating risk"*, (pp. xvii-xxix). Facet publishing, London, UK

Respondents:

R1: Head of the ESRB Secretariat (phone conversation)
R2: Employee at a National Financial Supervisory Authority in EU, country A (mail conversation)
R3: Employee at a National Financial Supervisory Authority in EU, country B (phone conversation)
R4: Employee at a National Financial Supervisory Authority in EU, country C (phone conversation)
R5: Employee at a National Company Registration Office (interview)
R6: Employee at a National Company Registration Office (telephone conversation)