Tidsskriftet Arkiv

VOLUM 8-1 2017

DOI: http://dx.doi.org/10.7577/ta.1959

ISSN 1891-8107

PEER REVIEWED ARTICLE

Ahmed Elragal and Tero Päivärinta

Opening Digital Archives and Collections with Emerging Data Analytics Technology: A Research Agenda

ABSTRACT

In the public sector, the EU legislation requires preservation and opening of increasing amounts of heterogeneous digital information that should be utilized by citizens and businesses. While technologies such as big data analytics (BDA) have emerged, opening of digital archives and collections at a large scale is in its infancy. Opening archives and collections involve also particular requirements for recognizing and managing issues of privacy and digital rights. As well, ensuring the sustainability of the opened materials and economical appraisal of digital materials for preservation require robust digital preservation practices. We need to proceed beyond the state-of-the-art in opening digital archives and collections through the means of emerging big data analytics and validating a novel concept for analytics which then enables delivering of knowledge for citizens and the society. We set out an agenda for using BDA as our strategy for research and enquiry and for demonstrating the benefit of BDA for opening digital archives by civil servants and for citizens. That will – eventually - transform the preservation practices, and delivery and use opportunities of public digital archives. Our research agenda suggests a framework integrating four domains of inquiry, analytics-enhanced appraisal, analytics-prepared preservation, analytics-enhanced opening, and analytics-enhanced use, for utilizing the BDA technologies in the domain of digital archives and collections. The suggested framework and research agenda identifies initially particular BDA technologies to be utilized in each of the four domains, and contributes by highlighting a need for an integrated "public understanding of big data" in the domain of digital preservation.

Keywords: digital preservation, digital archives, big data analytics, text mining, research agenda

1 Introduction

The European Union legislation (e.g., the PSI Directive 2003/98/EC) as well as national legislations (cf. Serra 2014) require increased opening of public digital archives and collections. The potential value for opening data in the public sector is expected to be tremendous. McKinsey estimated the annual global potential for producing economic value based on open data to be within USD 3-5 trillion in 2013 (Manyika et al. 2013). The European Commission estimated the aggregate economic impact from applications based on open government data across the EU27 economy to be €140 billion annually in 2012 (CapGemini 2013). Amounts of digital content continue to expand rapidly also in the public sector while human resources of information professionals to manage that data are rarely increasing. Data formats in government-owned archives and digital collections remain heterogeneous, sometimes becoming more complex along with technological developments. There is a reason to call the ever-increasing amounts of digital, heterogeneous, and public information reservoirs as big data (e.g., King et al. 2012). In 2012, Spanish annual report on open data estimated that so far only 10% of companies working on open data have been utilizing cultural digital collections (CapGemini 2013), while the recent Open Data Barometer (World Wide Web Foundation 2015) addressed the need for supporting city-level opening of government data together with sustained investment to support innovations on using open government data effectively. The World Bank only recently identified lack of robust digital preservation practices as one of the most significant hinders of sustained utilization of open data (Lemieux 2014).

2 Background & Challenges

Two reasons restrict human abilities to make best possible use of large digital data amounts preserved in the public sector: First, our human capabilities to collect, process, read, and comprehend data are limited, as data grows quickly in size and complexity. Second, computers – that can process large amounts of data far better than we can – still have difficulties in understanding the meaning of content, as natural language is inherently fuzzy and highly ambiguous. Advances in search-engine technology had an immense impact on our society, which shows the huge transformative potential of big data analytics (BDA). Anyhow, data search is only one of various BDA applications, and there is a lack of frameworks and ready-to-use technologies that support integrated knowledge extraction from digital archives and collections. The emerging field of BDA provides the ecosystem that is needed for their development. In particular, technologies for natural language processing, text mining, and information visualization can be applied to extract knowledge from public sector documents and other [open] data sources. Opening of digital archives and collections in the public sector involves also particular quality and security requirements on the data processing. These trends lead us to identify emerging challenges, which provide background and motivation for this research:

1) The public sector cannot feasibly preserve "everything" it produces. That is, both the sheer amount of digital content produced in the public sector and scarcity of human

resources to focus on content preservation, quality assurance, and publishing set significant limits on the preservation and open data activities (cf. Serra 2014). Hence, public organizations need to consider two significant challenges necessary to be scrutinized before digital archives and collections can even be considered to be opened:

- *How to appraise relevant digital content to be opened and preserved?*
- *How to process and prepare the digital data for opening and preservation in the age of scarce and expensive human resources available for digital preservation and information processing?*

2) The public sector cannot open "everything" it preserves. That is, opening of data involves several legal and institutional challenges, which require common policies and practical solutions for data processing (Gollins et al. 2014; de Rosnay & Janssen 2014). As well, the resources to publish and serve the audiences with large data quantities will anyhow be limited. This causes the following challenge in connection to the task of opening already preserved digital archives and collections:
   - *How to respond to the challenges and constraints of opening digital content in the public sector?*

3) Users of public digital archives and collections cannot utilize "everything" as such. That is, sheer opening of the data is not enough without making it easily accessible, comprehensible, or linked with other resources and usable for citizens, commercial information product builders and decision makers (cf. Sataslaatten 2014). Hence, yet another challenge takes place after the opening process:
   - *How to enhance access to and use of the opened digital content?*

Altogether, opening data from digital archives and collections represents new work-related challenges for the professional archivists and other people with readily scarce resources (cf. Anderson et al., 2014). The above-mentioned challenges of opening archives and collections are also tightly interrelated to each other. For example, metadata required for enhancing access to and utilization of open governmental information needs to be prepared for appraisal and preservation of the data (Sataslaatten 2014). However, reliance on metadata alone might be also risky e.g., in connection to ensuring non-disclosure of potentially sensitive archives (Gollins et al. 2014). Hence, the public sector information professionals need enhanced support for "everything" - that is, the whole process from appraisal through preservation and opening to enhancing use of opened digital archives and collections needs to harness potential of emerging information technologies. This leads to the research question addressed in this paper:

- *How to utilize potential of emerging big data analytics technology throughout the process of opening digital archives and collections in the public sector?*

3 Big Data Analytics (BDA)

It is difficult nowadays to open a popular publication and not run into a reference to data science, analytics, big data, or some combination thereof (Agarwal & Dhar, 2014). Big data are data whose scale, distribution, diversity, and velocity require the use of technical architectures, analytics, and tools in order to enable insights that reveal hidden knowledge and create value to business. Three main features characterize big data: volume, variety, and velocity (aka the three V's). The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data. Data volume is the primary attribute of big data (Chen & Guo, 2016). Big data can be quantified by size in terabytes or petabytes, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it is coming from a greater variety of sources than ever before, including logs, clickstreams, and social media – or, as in our case, digitally preserved archives. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as eXtensible Markup Language (XML) or Rich Site Summary (RSS) feeds. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume. Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites. Some researchers and organizations have discussed the addition of a fourth V, or veracity. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations (Elgendy & Elragal, 2014).

3.1 Emerging BDA technologies and techniques

The interest in BDA research is on the increase. Google's adoption of the MapReduce was definitely a catalyst, which has led to a lot of developments in the area of BDA. Further, the development and deployment of Apache Hadoop has also opened the doors for organizations to process extremely large datasets that has never been possible. BDA is the use of advanced techniques, mostly data mining and statistical, to find (hidden) patterns in (big) data. BDA is where advanced techniques operate on big data sets (Russom, 2011). The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems (Elgendy & Elragal, 2014). A significant amount of these techniques rely on commercial tools such as relational DBMS, data warehousing, Extract Transform Load (ETL), online analytical processing (OLAP), and business analytics tools. During the IEEE 2006 International Conference on Data Mining (ICDM), the top-ten data mining algorithms were defined based on expert nominations, citation counts, and a community survey. In order, those algorithms are: C4.5, k-means, SVM (support vector machine), Apriori, EM (expectation maximization), PageRank, AdaBoost, kNN (k-nearest neighbors), Naïve Bayes,

and CART. They cover classification, clustering, regression, association analysis, and network analysis. Actually, not only organizations and governments generate data; each and every one of us now is a data generator (McAfee & Brynjolfsson, 2012). We produce data using our mobile phones, social networks interactions, GPS, etc. Most of such data, however, is not structured in a way so as to be stored and/or processed in traditional DBMS. This calls for BDA techniques in order to make sense out of such data.

3.2 BDA Challenges

Researchers and practitioners alike face various types of challenges when using big data analytics for prediction, for instance, privacy and security of big data (Newell & Marabelli, 2015), platform scalability, integration, etc. However, for our purposes we only focus on key challenges that might hinder the ability of BDA to elucidate knowledge from digital archives.

It has been noticed recently that big data research may have suffered from the so-called "*streetlight effect*". That is, the tendency of researchers to study phenomena for which there exist plethora of data, instead of studying relevant problems. To explain, most of the experiments and data-analytic research is relying on data from biggest data-driven companies e.g., Facebook, Twitter, Google, LinkedIn and Amazon. Great percentage of such studies is focusing on the data made available for researchers by those companies, for internal purposes. That is, such data may be either biased towards solving those companies' problems, and not necessarily the grand problems. For instance, (Lotan, Graeff , Ananny, Gaffney, Pearce, & Boyd, 2011) showed that Twitter has become favorite BDA research destination. Such choice, of Twitter, by researchers is justified by its relatively high-level of accessibility and the relative openness of its API. Together, such two factors, have led to a substantial number of studies dealing with Twitter data. That relative ease of Twitter data collection and analytics has – amongst other factors - lead to the phenomena of streetlight research in BDA.

A researcher will face a streetlight problem, is government has been discretionary in opening selected archives for research. However, if the digital archives required for the research is opened, findings can become trustworthy. Despite the fact that big data, being passively created and continuously collected, has opened the door for plenty of research to be conducted, research needs to be formulated around important problems (Rai, 2016).

Researchers can access data from data-driven companies e.g., Twitter, Facebook, Google, etc. via API. Application Programming Interface (API) is a tool created for developers to interact with data producers. For instance, Twitter has created an open API allowing developers to source Twitter data. The advantage of the API is to promote external innovation, based on data. Offering data externally allows developers to create products, platforms, and interfaces without the need to expose the raw data. As a byproduct, Twitter has capitalized on this model by the acquisitions of different companies in 2012 mostly built around their open API. However, Twitter does not guarantee the delivery of the tweets that match search criteria by researchers,

unless for a paid service. Data providers like GNIP and DataSift, handle twitter paid services (aka firehose). The firehose consists of an agreement between researchers and distributor of the firehose e.g., GNIP on tweets the researcher should receive. As the data providers receive tweets they are pushed directly to the end user. The Twitter API is offered for free, yet the Twitter paid service removes a lot of the usage restrictions imposed by Twitter but comes at a fee, which not all researchers could afford. That fee represents what is known as "*data monetization*" for Twitter. Unambiguously, researchers need to delimit their scope based on the data available. The key issue here is to be aware of the limitations of the tools employed and to detail one's research approach accordingly.

In the context of digital archives, directives such as PSI in the EU help to reduce the impact of data monetization on BDA-enabled research. Likewise, open government provides another mechanism to make government data available without restrictions. Such directive and initiative help remove data-barriers and hence democratize the use of data.

4 BDA as a Process

Numerous authors have discussed the potential of BDA for information systems (IS) research, and recently Müller et al. (2016) have even provided guidelines for employing BDA in IS research. They conclude that "reflecting on the guidelines, we can observe that each phase of the research process requires a revised set of actions and abilities" (p. 11) and advocate a skill set change for IS researchers with stronger emphasis on developing skills for data preparation and the deployment of analytical tools and cross-instrumental evaluation criteria. We believe that similar developments will take place also in the field of digital preservation. Below is a brief description of the four phases to conduct BDA research.

4.1 Acquisition

BDA starts with acquiring the data through copying, streaming, etc. Such acquisition requires good understanding of the domain (often business context) as well as the data. Datasets, from which we source data, should be described in terms of: required data to be defined; background about the data; list of data sources; for each data source the method of acquisition or extraction; and reporting the problems encountered in data acquisition or extraction. One of the challenges associated with big data acquisition is: on one hand, there exist too much data, while on the other hand; all acquisition requires time, effort and resources.

4.2 Pre-processing

Pre-processing activities include: check keys, referential integrity, and domain consistency; identify missing attributes and blank fields; replacing missing values; data harmonization e.g., check different values that have similar meanings such as customer, client; check spelling of values; check for outliers. In result, pre-processing provides a description of the dataset including:

background (broad goals and plan for pre-processing); rationale for inclusion/exclusion of datasets; description of the pre-processing, including the actions that were necessary to address any data quality issues; detailed description of the resultant dataset, table by table and field by field; rationale for inclusion/exclusion of attributes; and the discoveries made during pre-processing and their potential implications for analytics.

4.3 Analytics

The central step in BDA is analytics during which data mining, machine learning, statistics and other techniques, or models, are chosen and applied on the data. For the implementation of a technique (or model) numerous number algorithms are available to be applied to any dataset. It is important on such step to describe: model assumptions, model description (e.g. rule-based models list the rules produced in addition to their accuracy and coverage), and results assessment (e.g. why a certain modeling technique and certain parameter settings led to good or bad results).

One efficient approach to follow in BDA research is to identify, early enough, what one is looking for. However, data scientists responsible for the analytics process are often not aware of this challenge and/or do not have the necessary knowledge to apply in that manner. Hence, it turns out that preferences of the data scientists and their education might drive the analytics part instead of the problem in hand, leading to insufficient knowledge discovery.
In particular, the selection of algorithm's parameters by the data scientist has a profound impact on analytics. A parameter is a value to fine-tune an algorithm. For any BDA tool e.g., RapidMiner, there are often a large number of parameters that can be adjusted. Listing the parameters and their chosen values, along with the rationale for the choice, is a key task. For instance, for the K-Means clustering algorithm, setting the number of k is a parameter. Too big k might not be useful for the decision maker and too little value for k as a parameter might not solve business problems. While empirical research stands on a solid foundation of measurement, data scientists tend to overlook the fact that algorithms parameter setting not only impacts analytics, but interpretation as well.

On the other hand, BDA provides us with higher accuracy prediction example techniques include: Support Vector Machine (SVM), Naïve Bayes Classifiers, topic modeling (see below), and Random Forests. BDA utilizes algorithms that are good in predicting future or unknown events.

4.3.1 Text Mining

In this subsection, we explore some of the analytics techniques that are very relevant to digital archives since its data is mostly in form of text, therefore we discuss text mining. Text mining is the process of deriving knowledge or quality information form text. Text mining is the process of deducing qualitative or unstructured pattern from text dataset. Text mining requires (text) documents to be pre-processed. Text pre-processing is a complex activity; hence, it is broken down into sub processes as explained below:

- *Tokenization*: tokenizing transforms each and every word or character in documents' text into a token;
- *Change case*: Change case transforms the case of all the tokens to upper case or lower case to avoid treating them like two tokens;
- *N-grams*: once tokens have the same case, then n-grams operator takes place and this operator merge different tokens together to form structures that have a meaning. To illustrate, assumingly we have 2 tokens: Big and Data. The n-gram operator will fuse them into "Big_Data" and that normally improves accuracy of later classification;
- *Stemming*: the process of stemming brings the word to its stem. That is, to make sure that the same words, which have various tenses, adjectives, nouns, etc. are considered the same. Also, this enhance classification accuracy;
- *Filter Stopwords*: the filter stopwords operator removes stopwords e.g., an, a, or, etc. This ensures that the words being matched together will be keywords not just simple stopwords;
- *Filter tokens*: the filter tokens operator takes the minimum and maximum length as a parameter.
- *Lemmatization*: Lemmatization refers to doing things properly with the use of a vocabulary and morphological analysis of words. It normally aiming at remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

Generally speaking, text mining techniques are used to discover the relation between text i.e., words, and this is achieved via multiple techniques e.g., clustering.

4.3.2 Topic Modeling

Topic modeling is a developing stage of statistical analysis of text. It is a mixture of ideas driven from many sciences like computer science, mathematics and cognitive science. Its progress is noticeable as it aids users understand and navigate large sets of unstructured data. Both Bayesian statistics and machine learning are used by topic models that result in benefits: - highlighting on thematic content of an unlabelled document; - predict the nature of future documents; - provide application-specific roadmaps through Bayesian statistics and machine learning. Many types of documents like emails; surveys and scientific abstracts have been the application of topic models. It is applied by looking for patterns of words used and interrelating documents with each other that show similar patterns. Topic models have appeared as a powerful technique for exploring and forming useful structure in the unstructured datasets.

4.4 Interpretation

Interpretation should relate analytical findings to the existing body of knowledge as well as industry practices and include reflection on certain business objectives, decision making, problem solving, etc. One of the significant problems in this step is the interpretation of 'quick & dirty' pattern discovery. The reason is mainly attributable to the fact that analytics can run easily and quickly by the data scientist, even via cloud. Given these opportunities, the pressure to reach outcomes often supersedes the genuine objective of the advancement of knowledge.

Another issue is the contradiction of predictive and explanatory power. Often, BDA provides us with higher accuracy prediction, but this accuracy comes at a cost. That is, most accurate algorithms such as SVM, Naïve Bayes Classifiers, topic modeling in text analytics, and Random Forests are not easily comprehensible by most of those who are supposed to consume their results, e.g., users and analysts of digital archives. In other words, BDA utilizes algorithms that are good in predicting future or unknown events and identifying patterns in data, but unable to provide easy-to-comprehend explanations.

5 BDA-enabled framework

Digital archives and collections in the public sector have the potential to become significantly more useful and usable for citizens, businesses and application developers with help of emerging data analytics technology. However, civil servants working as information professionals have scarce human resources to conduct digital preservation and open data deliveries in the world of ever-increasing amounts of digital data. Hence, also the appraisal and opening processes of archives and collections need to be supported with adequate data analytics technologies through understanding use better and for supporting the opening process with observance of legal constraints. The preservation task and its results, i.e., the archived information packages and metadata, need to be re-considered for enhancing utilization of data analytics. Altogether, innovative re-thinking towards analytics-enhanced appraisal, analytics-enhanced opening and use, and analytics-prepared preservation will form a dynamic service concept that needs altogether technological support to equip civil servants for better delivery of open digital archives and collections and to involve citizens, businesses, and application developers further in innovative ways to use them. See Figure 1.

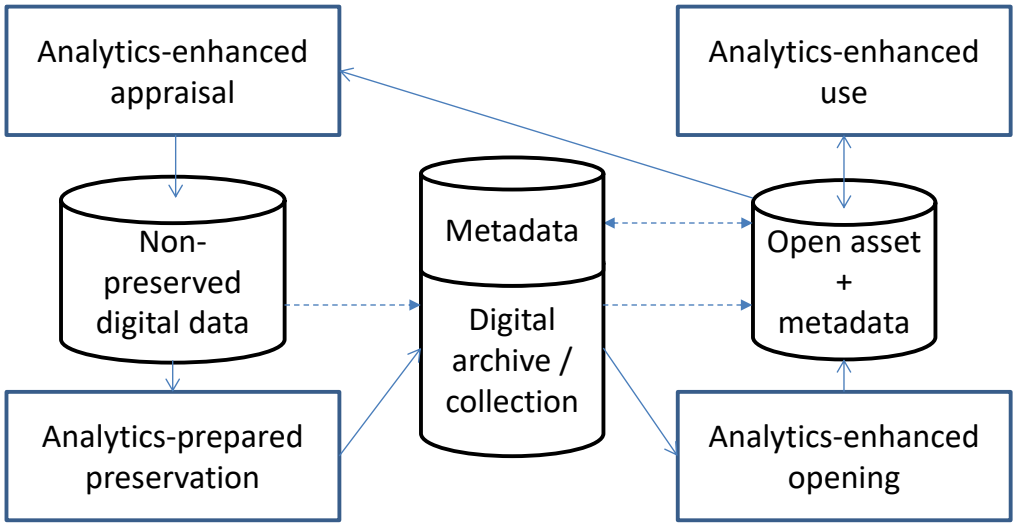

Figure 1 Four Concepts for Analytics-Enabled Digital Archives

Appraisal, preservation, opening, and innovative utilization of public digital archives and collections are challenging tasks even individually, especially in areas where the archived data is received and stored in heterogeneous formats and in the world of ever-increasing amounts of data production. This makes it essential to make the digital preservation and dissemination processes and the subsequent holistic service concept in the public sector dynamic, maximally automated and continuously improving. That is, the civil servants working as information professionals need to learn how to utilize data analytics to enhance relevance of appraised information, how to digitally preserve information so that its utilization can be maximally enhanced, how to adhere to legal constraints and issues adequately in relation to opening assets for dissemination, and how to boost usage opportunities for varying information customers to develop the service as a whole. Data visualization and interactive use opportunities provided by data analytics will as such provide useful means for users to access to the opened assets. Hence, the main constructs of our research come in a form of four interrelated concepts:

i. **Analytics-enhanced appraisal** aims at using big data analytics, especially Discovery Analytics, to support appraisal of digital records assets – which has previously been based mainly on human labour and expertise. Based on the opportunities provided by emerging technologies the concept will include novel appraisal methods and revise the traditional appraisal tasks and practices to be based upon discovered value of preserved records, records usefulness and observed risks of potential records misuse. That is, data analytics will play an important role as an enabler for appraisal to involve analysis of usage patterns of previously opened assets, comparing those to the materials potential for appraisal, and enhancing relevance of the preserved material and its potential usage. After the appraisal process, the preservation process will be informed about which non-preserved digital data (e.g., originating in government information systems or donators of data for cultural institutions) should be prepared for further preservation as a part of digital archives and collections and to be opened;

ii. **Analytics-prepared preservation process and data/metadata model** aims at developing submission and archival information packages (SIPs, AIPs)[1] and elaborated workflow model and solutions for digital preservation, which will prepare SIPs/AIPs and relevant metadata attached to them for further utilization of data analytics for opening and use of the materials. One important area is to utilize data analytics tools to support management of large datasets that needs to be restructured and assembled according to defined rules in the SIP preparation process. One potential cause for the need of restructuring of datasets could be based on the outcome of a sensitivity review process (cf. McDonald et al. 2014). Archival information packages need also to be prepared for potentially many-sided use of data analytics technologies that may need to gather information from them, sometimes even dynamically, for enabling new kinds of

[1] For the standard digital preservation terminology, see the Open Archival Information System (OAIS) specification (CCSDS 2004).

dissemination information packages to respond varying user needs who want to access to digital archives and collections from varying viewpoints;

iii. **Analytics-enhanced opening** is required for overcoming recognized legal, policy-related, and potential ethical challenges that may hinder opening of digitally preserved archives and information collections. Under this concept, such hindrances will first be observed, potential of emerging data analytics technologies to overcome such challenges will be reviewed, and selected ideas to utilize data analytics to overcome observed legal, policy-related and ethical hindrances of opening are to be demonstrated. While the traditional approach has been a human-conducted sensitivity review record-by-record (cf. McDonald et al. 2014), the concept of analytics-enhanced opening needs to go beyond plain human-conducted sensitivity reviews to innovate new approaches to creating opened dissemination information packages from the digitally preserved archival information packages that have been, in turn, produced through the process of analytics-prepared preservation. Data analytics can make different approaches to open digital archives possible: for example, potentially sensitive AIPs can be transformed to intermediary anonymized or pseudonymized resources to be opened in various ways later on, or less sensitive AIPs can be opened as such for further data analytics in connection to use. The alternative design choices will be scrutinized in more detail during the research;

iv. **Analytics-enhanced use** will provide significantly increased potential to utilize opened collections. Examples of development possibilities under this concept involve a scale from using simple visualization technologies on opened digital for citizen browsing and learning (e.g., in the context of schools and education) towards involving interactive content/text mining platforms by advanced users to form their own views and combinations even across collections of opened materials. The research will examine potential of combining established analytics-based use components with simple application development platforms to demonstrate potential for enhancing development of applications that can be based on opened collections and archives.

6 A Research Agenda

In this section, we provide a suggested research agenda to conduct research opening digital archives and collections with BDA. Our suggested research agenda is based on the BDA-enabled framework components: preservation; appraisal; use; and opening. We present opportunity for BDA associated with suggested research agenda in relation to each of those components as presented in the table below.

BDA-enabled framework	*Challenge [in brief]*	*Opportunity for BDA-enabled research*
BDA-enhanced appraisal	Using big data analytics to support appraisal of digital records assets – which has previously been based mainly on human labour and expertise	• *How to use BDA in order to support appraisal activities? e.g., the use of SVM in appraisal* • *How to use BDA in ranking and classifying assets? e.g., the use of K-Means Clustering to group similar assets* • *How to use BDA in order to project and predict the use of certain asset? e.g., the use of Random Forests in projections*
BDA-prepared preservation	Aims at developing submission and archival information packages and elaborated workflow model and solutions for digital preservation, which will prepare SIPs/AIPs and relevant metadata attached to them for further utilization of data analytics for opening and use of the materials	• *How could BDA deduce models in order to reduce the need to directly use digital archives [hence, enhance preservation]?* • *How could BDA facilitate preservations to minimize restrictions on access to digital archives?*
BDA-enhanced opening	Required for overcoming recognized legal, policy-related, and potential ethical challenges that may hinder opening of digitally preserved archives and information collections	• *How could BDA be used in order to capture and preserve security challenges?* • *How could BDA be used in order to address privacy and sensitivity concerns?* • *How could BDA be used in order to capture enacted roles and responsibilities integrated with/in digital archives?*
BDA-enhanced use	Provide potential to utilize opened collections from using simple visualization technologies towards involving interactive content/text mining platforms	• *How could BDA enable deep understanding of digital archives?* • *How to use text analytics in understanding digital archives?* • *How probabilistic topic modeling could shape our understanding of digital archives and hence open doors for new business models?*
Ecosystem related issues	• *In which way, can we assess the value delivered by BDA to digital archives and collections?* • *What kind of innovative services utilizing BDA in association with digital archives could be envisioned?* • *What changes does BDA enable with regard to operations of government agencies and their digital archives?* • *What are the design principles of BDA-based digital archives?* • *How to address the challenges facing BDA researchers – inevitably will also face digital archives researchers?*	

Table 1. Research Agenda

Recently, research on digital preservation has started to identify applications of data analytics in well-targeted research domains, such as tackling the preservation and metadata creation challenges of large content volumes (Schmidt et al., 2014) (related to analytics-enabled preservation), or varying solutions aiming at analytics-enhanced use, such as MERRA analytic services on climate data (Schnase et al., 2014), the Irish Record Linkage project on interlinking historical population registration data for enhanced use (Beyan et al, 2014), and data analytics for

legal records (Borden and Baron 2014). However, our four interlinked concepts for analytics-enabled digital archives form altogether a dynamic and integrated process to enhance automation and continuously improving preservation and opening of digital archives and collections. Hence, we argue that future research agendas on building platforms for utilizing BDA technology on digital preservation and archives should take, at least, these four dimensions into account altogether in order to cope with the increasing demands for automation and open data. Our work also suggests particular BDA techniques to be assigned to the appraisal, preservation, opening, and use solutions for digital archives. As such, the four concepts introduce the wide application opportunities of BDA technology and techniques to the field of digital preservation, beyond the previous prototypes of BDA-related preservation preparation and analytics focusing solely on the archive uses. Our work also suggests that the challenge of increasing "public understanding of big data" implications (cf. Michael and Lupton, 2016) in the field of digital preservation is still in its infancy to be tackled, while we would suggest the envisioned solution to be rather bold to connect BDA-enabled preservation to maximally automated appraisal, preservation, and opening which would even learn dynamically from the enhanced use patterns over time (cf. Figure 1 above).

7 Conclusion and Future Work

This research addresses the challenges of digital preservation, opening and delivery of digital archives and collections, through innovative utilization of emerging big data analytics technology. We suggest that BDA-enabled digital archives have the potential to enhance further innovations on the areas of appraisal, preservation, opening, and use of digital archives and collections. Our future work aims at reaching impacts on integrating adequate BDA technologies to support opening and use of digital archives and collections and to use them in an integrated manner with digital preservation and sensitivity review technologies under a service concept to involve an integrated view on appraisal, preservation, opening, and use. Based on such integration, our future research aims at impacting efficiency of the civil servants, who work with preserving and opening archives and collections in the public sector, effectiveness to be able to open materials which was not easy to open before and to open more relevant materials, and increased openness through possibilities to open new kinds of materials and making them more accessible and usable. Based on initial inquiries on benefits that can be reached from electronic archives in the public sector (cf. Päivärinta et al., 2014), we can, still, only imagine the huge scale of such public sector cost of storing information in closed, outdated systems instead of robust and standardized digital preservation. If more automated preservation practices that also would take the requirement to open the relevant resulting archives into account can be promoted, it would represent huge saving potential in the cost of preservation, let alone in the cost of, and benefits to be expected from, opening the preserved digital archives. Any achievement on the concept of enhanced appraisal will produce additional potential for efficiency and effectiveness gains, as the sheer amount of data currently residing in both active and closed systems is large, and the public sector needs to be also very selective on the collections with which to use their scarce human resources to open (or, even, to preserve beyond any minimum level legal compliance).

In the future, solutions on our framework will be prototyped with relevant public sector partners and their digital archives for streamlining and further enhancements.

References

Agarwal, R., & Dhar, V. (2014). Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information Systems Research, 25 (3), 443–448.

Anderson K., Engvall T., Klett E. (2014). Open Data through the Archive: A New Role for Archivists. In Arxius I Indústries Culturals, Girona, Oct 15th. http://www.girona.cat/web/ica2014/ponents/textos/id75.pdf.

Beyan, O., et al. (2014). Towards Linked Vital Registration data for Reconstituting Families and Creating Longitudinal Health Histories. KR4HC Workshop (in conjunction with KR 2014).

Borden, B.B., & Baron, J.R. (2014). Finding the Signal in the Noise: Information Governance, Analytics, and the Future of Legal Practice. Richmond Journal of Law & Technology 20(2), Article 7.

CapGemini (2013) The Open Data Economy. Unlocking Economic Value by Opening Government and Public Data. Capgemini Consulting Feb 7, 2013. https://www.capgemini.com/resource-file-access/resource/pdf/the_open_data_economy_unlocking_economic_value_by_opening_government_and_public_data.pdf

Chen, G., & Guo, X. (2016). Big Data Commerce. Information and Management.

De Rosnay, M. D., Janssen, K. (2014). Legal and Institutional Challenges for Opening Data Across Public Sectors: Towards Common Policy Solutions. Journal of Theoretical and Applied Electronic Commerce Research 9(3), 1-14.

Elgendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. The 14th Industrial Conference on Data Mining (ICDM). Petersburg: Springer-LNCS.

Gollins T., et al. (2014). On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records. PIR'14. Luly 6-11, Gold Coast, Australia.

King R., Schmidt R., Becker C., Schlarb S. (2012) SCAPE: Big Data Meets Digital Preservation. ERCIM news 89, 30-31.

Lemieux V. (2014) Why we're failing to get the most out of open data. World Economic Forum. AGENDA. https://agenda.weforum.org/2014/09/open-data-information-governance-quality/

Lotan, G., Graeff , E., Ananny, M., Gaffney, D., Pearce, I., & Boyd, D. (2011). The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions. International Journal of Communication.

Manyika et al. (2013). Open data: Unlocking innovation and performance with liquid information.

McAfee, A., & Brynjolfsson, E. (2012, October). Big Data: The Management Revolution. Harvard Business Review (HBR), 3-9.

McDonald, G., Macdonald, C., Ounis, I., Gollins, T. (2014) Towards a Classifier for Digital Sensitivity Review. In ECIR 2014, Springer-LNCS 8416, 500-506.

Michael, M., & Lupton, D. (2016). Toward a Manifesto for the 'Public Understanding of Big Data'. Public Understanding of Science 25(1), 104-116.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. Journal of Strategic Information Systems, 24, 3-14.

Päivärinta, T., Samuelsson, G., Jonsson, E., & Swensson, E. (2014). Nyttorealisering av FGS:er: Delprojekt 2. Project report. Riksarkivet, Stockholm.

Rai, A. (2016). Synergies Between Big Data and Theory. MIS Quarterly, 40 (2), iii-ix.

Russom, P. (2011). Big Data Analytics. TDWI, 4th Quarter, 1-38.

Sataaslaatten O.H. (2014). The Norwegian Noark Model requirements for EDRMS in the context of open government and access to governmental information. Records Management Journal 24(3), 189-204.

Schnase, J.L., et al. (2014). MERRA Analytic Services: Meeting the Big Data Challenges of Climate Science through Cloud-Enabled Climate Analytics-as-a-Service. Environment and Urban Systems, on-line pre-print. http://dx.doi.org/10.1016/j.compenvurbsys.2013.12.003

Serra L.E.C. (2014). The mapping, selecting and opening of data. The records management contribution to the Open Data project in Girona City Council. Records Management Journal 24(2), 87-98.

World Wide Web Foundation (2015). Open Data Barometer. Second Edition, January 2015. http://barometer.opendataresearch.org/

Ahmed Elragal (PhD, MBA, BSc) is an associate professor of information systems at Luleå University of Technology in Sweden. Prof. Elragal has over fifty research papers and articles published at international conferences and journals. He is a member of the editorial board of I & M and IJBIR journals. Prof. Elragal is the Associate Editor of the International Journal of Information Systems and Project Management (IJISPM). He is the co-author of Pearson's AWE of the MIS textbook (Laudon, Laudon, and Elragal). He is the winner of the 2010's international case study competition on "Business Intelligence", a prestigious international award. He has over 15 years of consulting experience, focused mainly on enterprise systems and business intelligence. He has helped different regional as well as multinationals organizations [including SAP, Teradata, Hyperone & Egypt Census Bureau] in the areas of enterprise systems, business intelligence, data mining, and big data [analytics].

Tero Päivärinta's research has lately focused on digital preservation and, especially, design science research with focus to integrate enterprise content management with long-term digital preservation services. Before his career at LTU, his research has focused significantly on enterprise content management in close collaboration with engineering/oil industry and the public sector alike. His PhD from the University of Jyväskylä, Finland (2001), focused on enterprise document management representing one of the first doctoral dissertations on the topic in its time. Tero has published more than 70-refereed articles in international information systems conferences and journals, including European Journal of Information Systems, Information Systems Journal, Information & Organization, Information & Software Technology, JITTA, Communications of the AIS, Scandinavian Journal of Information Systems, and Transforming Government. His current research interests include systems and software development practices, enterprise content management, preservation of digital information, and e-Government