Position papers of the Federated Conference on Computer Science and Information Systems, pp. 171–176
DOI: 10.15439/2017F549
ISSN 2300-5963 ACSIS, Vol. 13

Information Quality Challenges for the Preservation of Norwegian Public Sector Records

Markus Helfert
Dublin City University
School of Computing
Glasnevin,
Dublin, Ireland
Email: markus.helfert@dcu.ie

Petter Reinholdtsen
University Center for Information Technology, University of Oslo,
Oslo, Norway
Email:
petter.reinholdtsen@usit.uio.no

Thomas Sødring
Oslo and Akershus University
College of Applied Sciences, Oslo,
Norway
Email: Thomas.Sodring@hioa.no

***Abstract*— The digitalization from paper-based to electronic records management results in challenges to preserve material in an authentic form. This paper explores the role in ensuring the authenticity and usability of electronic records in a long term preservation perspective. The discussion is viewed from an information quality perspective that provides a suitable lens to the topic. We identified a number of challenges that government records face, stressing the issue at hand of how to maintain authenticity and usability over a long term perspective. Challenges results from issues around authenticity, usability, the user, volume and heterogeneity and particular the time dimension. We conclude that the fields of record keeping and long term preservation have some clear information quality issues that could benefit from a concerted approach by integrating information quality research into records management. To date this seems to be missing.**

I. Introduction

THE ongoing digitalization from paper-based records management to electronic records management challenges our ability to preserve material in an authentic form. Rubber stamps and signatures have been replaced with electronic signatures based on cryptographic principles and paper has been replaced by collaborative online word processing or reduced to semi-structured information stored in databases. This paradigmal shift away from paper potentially allows government to be more cost effective as manual labor intensive processes are replaced by software. However it also introduces new challenges for content preservation, as the material that is amassed no longer has a physical form, rather it is often reduced to binary data unreadable or accessible by humans. This results in a significant challenge for records management and digital content preservation. The reason for this can probably be summed up in the following statement "I can easily pick up and read a book that's 150 years old, but I cannot easily read the data stored on a tape that was connected to my Commodore Plus/4 in 1985". Technological obsolescence is an issue that challenges the ability to preserve electronic information.

This position paper focuses on two distinct professions working with government records, record keepers and archivists and is concerned with the process of extracting records from a relational database and depositing them with an archival institution. The results are an overview of our findings from a preliminary analysis of 12 years of government records from 1999 to 2012, for 5 various medium-sized Norwegian municipalities. The case-handling records were to be extracted a relational database adhering to the Noark standard. If we look at the digitisation process, there are many and various challenges that could be discussed. We have identified challenges emanating from an analysis of a number of databases containing electronic records when transferring records from the recordkeeping phase to the long term preservation phase. In this work we limit ourselves to challenges that became evident from our analysis, but there are other relevant challenges that could and should be identified and addressed. Another fact that complicates this discussion is that various countries have various approaches to record keeping and long term preservation, and as such the results may not be applicable to all countries.

The position we put forward for discussion is "*Information Quality has an important role in ensuring the authenticity and usability of electronic records in a long term preservation perspective*"

II. Background

The domain of government records is all records that are generated from government action and governments have a duty and obligation to preserve such records [Public Records Act 1958]. At the state level, national tax and infrastructure planning are prime examples, while at municipal level, planning, local healthcare, transportation, child protection services etc. can exemplify government

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094 to Lero (www.lero.ie)

records. Very often these two levels of government interact and this interaction also produces records which also need to be preserved. In addition modern societies with complex tax and welfare programs will naturally generate more records than countries without such programs.

Government records are collected and subjected to long term preservation for evidential, legal, fiscal, informational or historical purposes [Schellenberg] and for an archive, achieving evidentiary value for its collections means maintaining authenticity and usability with a perspective that spans hundreds of years. With this in mind, a new challenge arises about the role that information quality could play with regards to the preservation of electronic records for a long time period, say in a 1000 year perspective. The field of Information quality has typically been concerned with the recordkeeping phase of electronic records and little work has been done on information quality from a long term preservation perspective [Conway 2011]. We have identified the following as relevant challenges that should be addressed when considering Information Quality and preservation: authenticity, usability, user, volume, heterogeneity and most important time.

III. Challenges

A. Authenticity

One of the cardinal requirements for records is to maintain evidentiary value [Schellenberg]; to achieve this it must be possible to identify the authenticity of records. If a record is to be deemed authentic, it must be possible to prove that; the record is what it purports to be, that time elements (creation, dispatch, reception) are correct and involved individuals are identified correctly. Integrity is a property that is closely related to authenticity and often for a record to be deemed authentic it needs integrity. In this regard integrity can be defined as the degree of completeness of a record and whether or not we can prove record have been altered. Trust in records can be defined by whether or not we deem them as authentic. With paper, it is relatively easy to identify authenticity. Money is a very good example of this, where there are mechanisms in place so people know they can trust a monetary note as authentic.

How do we bring such trust mechanisms into play with electronic records? Typically, hashing [REF] is used on records and documents to determine if a record has been accidentally altered. That is a relatively, simple cheap and sufficient technique to determine integrity against bit rot or accidental changes to a record. Hash algorithms are also suitable for long term preservation as they are well documented and many implementations exist. It will not protect against willful modification of records, where both the record and the hash is changed. As computers become faster, it might become computationally feasible to calculate how to change records and documents in a way that do not change the hash value. Hashing may solve part of the integrity issue but does not necessarily solve the authenticity issue. Hashing can solve authenticity if there are mechanisms in place to ensure we can trust the hash and original records. Documented processes, third-party logging of values, writing hash values to write-once-read-many media, trusted timestamping or writing hash values to a public blockchain are mechanisms to lift hashing from a integrity to a authenticity mechanism. Other approaches to authenticity include the use of public key infrastructure. Here an entity signs a hash of a record with their private key and their public key can be used to verify the signed record have not been changed. PKI may poses a particular challenge to preservation if the public key required to verify a signature is no longer available, or if the encryption mechanism has been broken to a point where it is computationally possible to create fake signatures. The public key is well know at the time it is used so the public key can easily be stored in relation to the record, but is only useful as long as the verification method is well known.

While the above points relate to individual records, dealing with large collections of records is also a challenge. Electronic records are typically stored in relational databases and it can be a challenge to preserve these over time. Just extracting database data as official records is a challenge and a natural question is quickly raised, How can we trust large database extractions? At the simplest level, one could create a database extraction as a backup and calculate a hash value for the extraction. This approach is common, but also naive. Who produced the extraction and how? Have all schemas been extracted, have the correct schemas been extracted?

Assuming authenticity without an understanding of usability is naive and dangerous and should be considered an information quality challenge.

B. Usability

In order to achieve usability, we must maintain storage, readability and understandability. If we are unable to store the information, we have lost it. Sometimes problems relating to storage are as simple as the fact that the people involved are not aware the data in a database is to be preserved and assume that once it is no longer in active use it can be deleted. When it comes to readability, an issue can be that a database backup extraction is seen as valid archival object. This can be problematic as the software to read the contents may fall away. This is a particular issue with databases from the eighties and early nineties. We know today that this issue is real and there are many examples of digital content that have been lost [OAG]

because we are simply no longer able to read the information. The municipality of Oslo for example has over 666 [OAG] various systems running on various database platforms. Preserving all of these databases is a challenge.

The SIARD file format (Software Independent Archiving of Relational Databases) is an interesting approach to this problem and solves the readability issue by converting the data in a relational database to a similar structure in XML. This data can then be imported back to a database for re-use, but also can act as a preservation object.

However, using relational database extractions as a preservation object has run foul of data privacy laws in Norway. The reason for this is that if it is not possible to identify records in the extraction, then it is not possible to be in compliance with data retention and disposal laws. A similar issue with readability is related to the use of older document file-formats. In particular document formats from the 80s that can no longer be interpreted and displayed by software is a challenge. While this issue is often over exaggerated as being a unsolvable problem, the authors did run into particular problems converting Lotus WordPro (lwp) files to PDF/A. The authors also had difficulty dealing with a variant of WordPerfect files stored in a database from eighties where the original frontend system likely had integrated fields for the automatic generation of documents and accurate reproduction of the document was not possible without the front-end system. This issue is one that can have a negative impact, not just on usability, but can also increase preservation costs significantly. The archive may have to preserve multiple versions of a document, instead of just an archive (suitable for long term preservation) version. The original version must be kept and examined when a user requests access to its contents. However at some point in the future, the tools to access the underlying data in such a document may no longer be available. In the case of the lwp files we examined, we were able to open them with a text editor, and even though there was some binary information there, a lot of the text was retrievable.

When it comes to understandability, a difficult issue to deal with is what is known as semantic drift. This is a result of the material being 'frozen' at the time it is archived, but society and the language we use evolves, slowly, over time. Over longer periods of time, this can have a major impact on understandability. If we consider records written in the 16 century in old English we know that the language used since then has evolved and this challenges both readability and understandability. It may be difficult for a layperson to read documents written in gothic script and even if you could read such documents you may not be in a position to understand the contents as the English language has evolved.

It is difficult to imagine the effect that globalization will have on languages and how many of today's languages will still be around in 500 years, but this is something an archive must be aware of. But even within a language, there can be many dialects and sayings that will ultimately be lost. A major information quality challenge is how records survive the test of time and be available for users.

C. The user

When we talk about the notion of a user, we should picture the grandchildren of the grandchildren of the grandchildren of our children. That is the perspective we need to have. Within information quality, the user is often a guiding factor for research [Wang and Strong 1996], but for government records, in a preservation perspective, the idea of a user is sometimes secondary as the material is often preserved because of law, not because a user wants to interact with it. The OAIS model [CCSDS] argues the need for a clear definition of a designated community must be defined for archival collections; however, we see that in practice, the designated community is often an afterthought and as such the notion of a user is very open. Who will the user of electronic records in fifty years or five hundred years be? Humans will most likely access the material for personal and research reasons, but also artificial intelligence stemming from research into big data will probably be developed to create an understanding on why society evolved the way it did.

A natural question to ask is whether or not we are capturing enough information for future needs and to identify the level of quality of the information we are capturing. When records were paper based, it was easy to identify them, read them and process related information like comments written on paper. When records became electronic, we lost the ability to integrate a "human aspect", the handwritten note, or the extra related piece of information that has no place within an electronic system. It is likely that we are simply throwing information away because systems became digital. But we do not know for sure.

Collecting and centralizing large amounts of records related to a user can pose sensitivity challenges. For example, a person requesting a copy of a school diploma may find it uncomfortable when the archivist searches through their records and sees that the person has been sexually abused by a teacher. Another examples is where a user comes across information that might damage the psychologically. Perhaps a psychiatrist has asked for information not to be disclosed as the person is liable to self-injury if the person discovers such information. There is a need to balance access to information both in terms of

today's users but also users in the future. Relevant examples here can be seen in the release of archival records relating to criminal proceedings dating back one hundred years. Some family members report embarrassment when they find out their great grandmother was arrested for prostitution. Similarly, in 2011 nude pictures of the arctic explorer Fridtjof Nansen were released. He had taken pictures of himself that he sent to his wife. These two examples show that even though archive material is static, the story about the material evolves and can evoke feelings long into the future.

D. Volume and heterogeneity

The heterogeneity of data sources that government both create and use mean that archival institutions need to understand database structures of potentially thousands of systems. With the current trend of big data [Laney 2001] the challenges can be expected to increase [Haug and Arlbjørn 2010]. The municipality of Oslo, for example, has 666 various systems [OAG] containing digital information about citizens that should be preserved. Within these systems you will find a variety of databases and document formats that are no longer in daily use. Compounding this issue is a rapid technological evolution that gradually introduces obsolescence over time as volume continually increases. From our studies, we see that recordkeeping institutions typically have a focused time-frame over records. This act likes a window into the records and covers mostly all on-going cases. Case files from years back are seen as more archival in nature and data quality issues are not that important. There appears to be a need to limit the view of records, to keep the volume down. The problem with volume comes to light when you need to extract the records after 12 years. If you have 12 years of problems, incorrect data entry, bad system design, the you will have problems with preservation. These problems need to be identified and resolved.

Heterogeneity of material is also a matter for concern. The more the heterogeneity in file formats and database structures, the more difficult and costly preservation becomes. It becomes difficult to create a coherent understandable extraction, but also costly for the archive institution to preserve. Many archives have guidelines with regards to what file-formats they will accept and it may be difficult for the records creator to be in compliance. If they for example have 2 million documents in 20 different file formats and versions of file formats, it is difficult to guarantee the conversion process. Here heterogeneity and volume cross each other creating an additional challenge.

E. Time

Time is a challenge not only because of the long period of time records are to be preserved, but also because it results in changes in technology and society and these have an effect on how we create, store and manage records. Time flows in one direction, nothing can stop that and if we do not address this challenge, we are creating digital mess that will be difficult to clean up. The term "technical debt" is often used to highlight the fact that an organization has hidden costs when dealing with the preservation of records. In the same way some argue that records are strategically important to an organization and should be managed as an asset, the mismanagement of records should be seen as a liability.

When it comes to preservation, time is both a challenge as well as a factor that compounds the other challenges. Over time, heterogeneity in the material and volume naturally increases along with technological evolution. Heterogeneity over time must be reduced by standardisation. We see this today with the use of the migration strategy for long-term preservation. As this is a cost issue, it is likely the archives will be forced to increase homogeneity.

The user and user expectations follows a similar trend with time, the technological evolution has given a new rise to user expectations of what an archive should deliver, today's generation will be impatient and have little understanding that archives cannot simply publish information. The user of tomorrow is likely to be based on big data /AI. Perhaps such algorithms will have a higher tolerance to bad information quality than humans do.

When it comes to usability, time is the very essence of semantic drift, but also a factor in the technological evolution that results in technological obsolescence.

At face value it would appear likely that authenticity mechanisms, like hashing, will not be affected by time. The hash, the data and the algorithm are all constant so there is little chance errors could occur. However bit-rot could be an issue on magnetic media, as a single flipped bit due to deteriorating media will cause a negative outcome when undertaking a hash check. Technology advancements might see concerted efforts to hack documents assumed safe by hashing. In much the same way bitcoin mining today attempts to create hashes, which follow a given pattern, we might find that future archive documents are subjected to attacks where the content of documents are replaced by malicious content with an equivalent hash. This could achievable by inserting dummy non-visible data into e.g. a PDF file and mining the dummy data until the file hash matches the hash of the original file. While such an attack vector is unlikely with todays technology, as technology advances, the possibility may increase. Multiple hashes, using various algorithms, for each document is an easy way to reduce this attack vector.

The same argument could be made with PKI, that mining could be a potential problem, even though it is not one today. It is possible to store the original data, the public key that verifies the data and the algorithm to that

undertakes the verification process, in the same way that hashing is used today. In Norway, we know that this is not practice.

Almost paradoxically, one can observe that, time is an issue that can have a negative impact on long term preservation.

IV. Discussion and Counterargument

The position we argue is that "Information Quality has an important role in ensuring the authenticity and usability of electronic records in a long term preservation perspective".

Information Quality is a mature and proven research field that has clearly made inroads to the field of record keeping. Its role in long term preservation is unclear though. The fields of recordkeeping and long term preservation are sometimes understood to be distinct, and that records exist in their own phases [Cunningham]. We believe this is the wrong view to have for records. Rather than solely focusing on issues within a single phase of a record, we should make information quality an overarching goal between the various phases. We need to understand the individual information quality needs during both recordkeeping and preservation. However when it comes to preservation, we can only inherit the inherent quality of the material and have little room to change anything. Fixing poor quality from the preservation phase is often so expensive, so fixing it becomes practically impossible, and as such any solution to preservation information quality must have its roots in the record keeping phase. However not all record keepers see preservation as their responsibility, so achieving acceptance for preservation information quality can be difficult.

To argue the other side of the position seems counter intuitive, that information quality has no role in ensuring authenticity and usability of electronic records. Of course it does, but approaching this from an information quality perspective may not be the only valid approach. A lot of research has been carried out in various disciplines that the archival profession has successfully used to push their own professional requirements. From the perspective of long term preservation, there is a need to bridge the gap between record keeping and long term preservation and using a formal information quality approach to bridge the two is worth exploring, We argue the need to pursue a more holistic approach to recordkeeping and long term preservation that finds its roots in the field of information quality.

We discussed the notion of a window into the electronic records. Such a window is a view of the records, where individual record are clearly defined and relevant information quality measurements can be readily available. The first course of action must be to figure out what such window would look like. How do we create a window into what is arguably a very dynamic, distributed and ever changing architecture. Initial experimentation would suggest such a window must become larger than a time span of weeks to months and cover all records, or the window must become very narrow and focused and only move when high information quality is achieved.

V. Conclusion

We identified a number of challenges that government records face when dealing with the process of records going from active use to long term preservation, where the issue at hand really is how to maintain authenticity and usability over a long term perspective. Our case is guided by some preliminary work on information quality on government records and we see the need for more research on this topic, that there is no one-size fits-all solution and the archival profession must aim to achieve understandability and just readability of records.

The challenges described above are cross disciplinary and fall within a number of disciplines. Some may argue that such issues are resolved, and perhaps at an abstract level they have been discussed within an academic context, but to the best of our knowledge there have been no studies on the information quality issue when looking at the process of extracting records from a record keeping system and transferring the records to an archive that cover the information quality requirements of the archive. The fields of record keeping and long term preservation have some clear information quality issues that could benefit from a concerted approach by integrating information quality into their respective fields and to push information quality as an overarching issue that ties the two fields together.

Acknowledgment

This work was supported by the Business Informatics Group at Dublin City University and in part, by Science Foundation Ireland grant 13/RC/2094 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Irish Software Research Centre (www.lero.ie).

References

[1] Arnon Rosenthal, Len Seligman, and Scott Renner. 2004. From semantic integration to semantics management Case studies and a way forward. ACM SIGMOD Rec. 33, 4, 44–50.

[2] Cinzia Cappiello, Francalanci, C. and Pernici, B., 2004. Time-related factors of data quality in multi-channel information systems, Journal of Management Information Systems, 20(3), pp. 71-91.

[3] Mouzhi Ge, Helfert M, 2008. Data and Information Quality Assessment in Information Manufacturing Systems Business Information Systems,

Volume 7, Lecture Notes in Business Information Processing (2008), pp 380-389.
[4] Anders Haug, Arlbjørn J.S. 2010. Barriers to master data quality, Journal of Enterprise Information Management. Vol. 24 No. 3, pp. 288-303
[5] CCSDS Secretariat, Space Communications and Navigation Office, 7L70. (2012). Reference Model for an Open Archival Information System (OAIS). Washington, D.C: UNT Digital Library.
[6] Cunningham, A. (2008). Digital Curation/Digital Archiving: A View from the National Archives of Australila. The American Archivist, 71(2), 530-543
[7] Dimitry Karagiannis, Mayr H., Mylopoulos J, 2016. Domain-Specific Conceptual Modeling, Springer.
[8] Doug Laney (2001), '3D Data Management: Controlling Data Volume, Velocity, and Variety' , Technical report, META Group .
[9] OAG. Office of the Auditor General of Norway: investigation of the efforts to secure and make accessible archives in the municipal sector. Oslo: Riksrevisjonen 2010. Document 3:13 (2009-2010)
[10] Public Records Act 1958. Available at vhttp://www.legislation.gov.uk/ukpga/Eliz2/6-7/51/contents (Accessed: 30 August 2016).
[11] Schellenberg, T. R. (1956) The Appraisal of Modern Public Records
[12] Stuart E. Madnick, Richard Y. Wang, Yang W. Lee, and Hongwei Zhu. 2009. Overview and framework for data and information quality research. ACM J. Data Informat. Qual. 1, 1, Article #2.
[13] Richard Y. Wang, 1998), A product perspective on total data quality management, Communications of the ACM, 41(2), pp. 58-65.
[14] Richard Y. Wang, and Strong, D.M. (1996), Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, 12(4), pp. 5-34.
[15] Conway, Paul: Archival quality and long-term preservation: a research framework for validating the usefulness of digital surrogates Archival Science, November 2011, Volume 11, Issue 3, pp 293–309