The migration and preservation of six Norwegian municipality record-keeping systems – lessons learned

Thomas Soedring[1], Pia Borlund[2], Markus Helfert[3]

[1] Department of Archivistics, Library and Information Science, Oslo Metropolitan University, Postboks 4 St. Olavs plass, Oslo, N-0130, Norway. Phone: +47 67 23 82 87, E-mail: Thomas.Sodring@oslomet.no [corresponding author]

[2] Department of Archivistics, Library and Information Science, Oslo Metropolitan University, Postboks 4 St. Olavs plass, Oslo, N-0130, Norway. Phone: +47 67 23 52 50, E-mail: Pia.Borlund@oslomet.no

[3] School of Business, Maynooth University, Maynooth, Co. Kildare, Ireland. Phone: +353 7008727, E-mail: Markus.Helfert@mu.ie

This is the peer reviewed version of the following article: Sødring, T, Borlund, P, Helfert, M. The migration and preservation of six Norwegian municipality record-keeping systems: Lessons learned. J Assoc Inf Sci Technol. 2020; 1– 11, which has been published in final form at DOI: https://doi.org/10.1002/asi.24438.
This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Abstract

This paper presents a rare insight into the migration of municipality record-keeping databases. The migration of a database for preservation purposes poses several challenges. In particular, our findings show that relevant issues are file-format heterogeneity, collection volume, time and database structure evolution, and deviation from the governing standard. This paper presents and discusses how such issues interfere with an organization's ability to undertake a migration, for preservation purposes, of records from a relational database. The case-study at hand concerns six Norwegian municipality record-keeping databases covering a period from 1999 to 2012. The findings are presented with a discussion on how these issues manifest themselves as a problem for long term preservation. The results discussed here may help an organization and Information Systems (IS) manager to establish a best practice when undertaking a migration project and enable them to avoid some of the pitfalls that were discovered during this project.

Keywords: migration, databases, record-keeping, data preservation, archival systems maintenance, legacy systems, data analysis

Introduction

Migration of records is a task that always carries a certain level of suspense associated with it, as it can be difficult to determine whether a migration process has been undertaken successfully or not. A migration is defined in ISO 15489 (ISO 15489:2016, 2016, p. 9) as the "act of moving records from one system to another, while maintaining the records' authenticity, integrity, reliability, and usability". A migration is commonly carried out when transferring records from one system to another, typically as a result of a system upgrade. Still, a migration can also be the conversion of documents from one format to another. There may also be a need, however, to undertake a migration for preservation purposes, e.g., when a system is no longer in active use, and there are legal requirements to retain the records. This work provides a unique insight into relevant issues that are problematic when undertaking a migration for preservation purposes and is based on a case study of a preservation migration process of the databases of six Norwegian municipalities with associated document collections (consisting of over 850 000 documents). The databases were developed and maintained to comply with a Norwegian record keeping standard called Noark[1] and allow us to study what consequences, if any, a formalized approach to record-keeping may have on long term preservation of government records.

This is the first time an empirical study reports on the extraction of data from a legacy system that shows the consequences of a system deviating from the governing standard. Hence the paper makes three significant contributions: 1) it reports empirical data from a migration project, 2) provides support and insight for an IS manager regarding migration, and 3) offers empirically-based recommendations on the migration and management of databases. The remaining paper is structured as follows: First, the record-keeping traditions in Norway are described with particular attention to the Noark standard. Then related research is reviewed before the theoretical framework is presented. After that, the data collection and the applied migration approach is described. Then the findings are detailed before the lessons learned are discussed. Finally, the paper wraps up with a conclusion.

Record keeping traditions in Norway

Norway is in a unique situation as it mandates the use of a standard for governmental and public administration record keeping and preservation, offering a standardized life-cycle approach to the handling of such records. This standard is called Noark, translating to Norwegian Archive standard, and its roots can be traced back to 1984. The use of Noark is enshrined as a

regulation in law[2], mandating government agencies to capture and preserve records in a standardized way. A Noark record-keeping system, according to Hagen Sataslåtten (2017, p. 9), is both a correspondence archive for public administration, as well as a documentation archive of the public administration in their function as executive authority. Another important aspect of Noark is that it serves as a source of records for freedom of information (FoI) requests (Hagen Sataslåtten, 2017, p. 9). Having a formalized approach to the record-keeping process for government records exposes records to the public, thus enabling easier access for citizens when undertaking formal FoI requests. The Noark standard has undergone multiple iterations over the years to keep up with the persistent evolution of information technology. The work detailed here concerns itself with databases defined per the fourth version of the Noark standard that was relevant between 1999 and 2008.

Related research

This contribution is placed within the areas of 1) record-keeping, 2) long term preservation, and 3) migration. The work presented here is a cross-disciplinary approach reflecting all three of the above areas. It offers a unique and rare insight into practical issues when undertaking a migration of record-keeping databases for preservation purposes. These areas are mature research areas with substantial contributions, with both reflections of practice (Duranti, 2005; Lorie, 2001; Ross, 2012), and literature from more general perspectives (Lin et al, 2003). Some notable ISO standards also cover these areas: Records management (ISO 15489:2016, 2016), Open archival information system (ISO ISO14721:2012, 2012), Digital records conversion and migration process (ISO 13008:2012, 2012), Principles and functional requirements for records in electronic office environments (ISO 16175-2:2011, 2011), (ISO 16175-3:2010, 2010), (ISO 16175-1:2010, 2010). The areas of record-keeping and long term preservation are interrelated (McKemmish, 2001), and approaches and practices between countries can vary significantly (Motsaathebe & Mnjama, 2009). Norway's governmental tradition, as described in the previous section, is based on a record-keeping phase followed by a preservation phase, where records are migrated and converted to a format suitable for long term preservation. Database migration is a broad area of research covering issues from migration between database models (Bisbal et al., 1999; Brodie & Stonebraker, 1995) to database conversions (Maatuk et al., 2008). The migration project presented here is specific to a particular context, while a lot of the existing literature is more general. Maatuk et al. (2008), for example,

discuss the problem of mapping between database structures from a high-level point of view, while Martens et al. (2018) focus on continued access to data between system migrations, rather than extraction for preservation purposes. High-level approaches, e.g., mapping databases to XML (Rahman et al., 2012) or other database models, do not provide that much insight as the approaches are too generic to provide concrete suggestions to solve the problem at hand. Another distinction that is worth making is that a lot of the related research here is about database migration, while this work concerns itself with record migration bound to a particular record-keeping standard. This gives the migration issue a different dimension when comparing it to existing related research. Lin et al. (2003), for example, discuss general issues related to volume and authenticity but have an institutional perspective on how to preserve objects that have been migrated, while this work is concerned with the previous stage, that is how to migrate records from a database with some underlying problems. The work presented here also covers document migration. Suri and El-Saad (2017) provide a qualitative and quantitative overview of the types of errors that can occur during the migration of documents to PDF/A and can act as an interesting backdrop to compare against. The work described here can also be seen as an enterprise architecture issue. Becker et al. (2011) propose an approach for preservation at the enterprise architecture level to ensure a coherent and consistent unified high-level view to help control the complexity of the preservation problem in a heterogeneous IT architecture. Their approach, however, has limited transferable value to this context. Finally, Lübeck et al. (2003) detail the migration process of a large volume (300TB) of data from physics experiments; however, they are more concerned with performance issues relating to software and hardware than matters relating to records. The present work positions itself within an inter-disciplinary arena of research, but as the literature review reveals a lot of semi-related work exists, but no previous work shares the focus or compliments the approach of our work.

Theoretical framework

The work presented here relates to a legacy system at its end of life, where a particular focus is applied to the overlap between the record-keeping and preservation phases of the record life-cycle. As such, the theoretical framework for this work has its roots in the handling of data from legacy systems. Bisbal et al. (1999, p. 2) note that the "lack of documentation and understanding of system" is one of many issues dealing with legacy systems. The work here is in a similar situation, as no documentation was available. One of the solutions for dealing with

legacy systems is migration, where data is mapped from one system to another (Bisbal et al., 1999, p. 5). This work takes a similar approach to migration, mapping data from a legacy database structure to a structure appropriate for solving the problem at hand (a predefined reference structure). The mapping from the original structure to an intermediate structure, where issues relating to standards compliance are corrected, is similar to the proposed migration tool by (Mellor et al., 2002). While their tool is a general approach to handling the migration of numerous heterogeneous digital objects, it can also serve as a guide to solving any database mapping issues. The strategy here also reflects the advice given in ISO 13008:2012 (2012), that covers the conversion and migration of records, and note that it is crucial to focus on maintaining existing relationships between objects. Our work analyses relationships between entities in a relational database looking for various types of relationships (e.g., aggregation, structural, functional, etc.) and ensure that they are maintained or strengthened where possible. ISO 13008:2012 (2012) also argue the importance of working on a copy of the records and documenting any information that is lost during migration. This advice is followed.

Data collection

The reported migration project is based on a technology claiming compliance with version 4 of the Noark standard. The Noark 4 standard consists of a detailed domain model and a recommended implementation of the domain model in a relational database (Sirevåg, 2014, p. 22). The standard details a list of 95 tables that should be in place for the database of a complete system. However, not all functionality is mandatory, so there is an expectation of variations in database models. The migration project is based on the record-keeping databases and document collections from six different Norwegian municipalities. The municipalities had purchased and deployed the same instance of a particular record- keeping software, but maintained their own databases. The systems were hosted by a shared IT-department and had a similar setup.

The average population count of the municipalities was approximately 3 700, while the population count of the largest municipality was just under 5 000. The record-keeping technology that sat on top of the databases was relatively old, dating back to mid-nineties. As such, the systems and databases have undergone multiple development iterations. The technology in question was in production for the six municipalities from 1999 to 2012, resulting in 13 years of record creation. It was noted that the software vendor claimed they could extract the records using a migration tool that was under development, but the municipalities would

have to carry a heavy burden of the development costs. As such, for all intents and purposes, the tool did not exist. There were 157 211 case files across the six databases, with the largest municipality accruing over 42 000 case files spread across the 13 years, while the smallest municipality had 14 500 case files. Each case file averaged roughly six registered documents per case file. The largest case file had more than 200 documents registered. This was a case file detailing a job application process for a particularly attractive position in the municipality. The document collection for the largest municipality was just over 22 GB in size. The largest file within this collection was a 235 MB PDF-document, which was a building application for the development of the town center.

Table 1 shows that there were 29 different registered file types, a fact that reflects how the system has been used and what users can expect to find there in the future. A general recommendation in public record-keeping is that records should periodically be deposited with an archive institution to ensure the municipality complies with its archival obligations. Given that, the municipalities had accumulated 13 years of records, and it became a pressing issue to preserve the material for future generations.

Table 1

A list of the count instances of various file formats across the six databases and their preservation equivalent. In some instances, number counts are rounded

Original file format	Preservation file format	Count
BMP	JPEG	100
CSS	N/A	1
DOC	PDF/A	327 000
DOCM	PDF/A	12
DOCX	PDF/A	1 360
DWF/DWG/DXF	PDF/A	33
EXE	N/A	5
GIF	JPEG	1 227
PDF	PDF/A	342 000
HTML	PDF/A	19 000
JPEG	JPEG	4 100
LWP	PDF/A	1 431
MOV	MPEG2	1
ODT	PDF/A	25
PPT	PDF/A	280
PPTS	PDF/A	1
PPTX	PDF/A	30

Original file format	Preservation file format	Count
PNG	JPEG	600
RTF	PDF/A	395
TIFF	TIFF	495
TXT	TXT	42 461
XLS	PDF/A	958
XLSM	PDF/A	1
XLSX	PDF/A	110
XPS	PDF/A	1
XML	XML	4 410
ZIP	N/A	64

Each database consisted of over 100 various tables. Within a database, it was possible to identify 70 tables that were relevant to the Noark 4 standard. The problem here was that, while the technology claimed compliance with the Noark 4 standard, it did not employ a database structure easily identifiable with the standard and would require extensive mapping of tables and columns to transform the database to the reference structure. The Noark 4 standard consists of a set of 95 tables. Some of the missing tables were Noark 4 functionality that was not implemented in the system, but the system also contained additional records that were not covered by the Noark standard.

Applied migration approach

Given the technical challenges, our expertise in record-keeping structures and migration processes was requested. We approached the migration project on an ad-hoc basis, which is a cardinal mistake when undertaking a migration. It is easy to be critical in hindsight that no detailed planning procedure was undertaken, where the work was better structured, but the migration project had no clear starting point. There were so many different tasks to be conducted. The first step was an analysis of the database and the Noark 4 standard, and it focused solely on solving the problem at hand from a technical perspective. The goal was to decipher the database to migrate the data for preservation purposes. The approach was very much investigative and ad-hoc in nature, and a proper understanding of the project did not emerge until about a year into the project. There was neither system nor technical documentation available, but access to user manuals and the running system was possible for a limited time as the system reached its end-of-life. Most of the time was dedicated to analyzing the original database structure, and mapping this database structure to the reference Noark 4 database structure, copying over records, fixing

any problems that were present. After this work was finalized, records were exported to their preservation format in XML, and the remaining project effort was concerned with migrating documents to a file format suitable for long term preservation.

Findings

A summary of the main findings show that the following issues were the main hinders in the migration project:

- Database evolution
- Database issues
- Deviation from the standard
- Incorrect system use
- Missing data
- Automated document migration

The findings point to a general requirement of ensuring compliance with a governing standard, where applicable and continual analysis of databases and document collections to ensure migration at some stage in the future will be possible.

Database evolution

Database evolution can be a challenge to migration as relationships between tables and columns can change. Data can be stored in different places during various periods, and such changes make it challenging to automate the migration process, especially when documentation is unavailable. In the databases that were investigated, multiple changes in database structure were observed. A significant upgrade to the record-keeping system in 2006 saw the introduction of multiple new tables. This was visible as data in these tables had the earliest date recorded as being in 2006. It was also discovered that records in other tables could take on a new form after 2006 when compared to how they were stored earlier. This can be exemplified in the way that comments created by case handlers, and record-keepers were stored. Such records should have been stored in a particular table called "comments" but were stored in multiple different tables. An assumption was made that the system previously interpreted comments as being in a one-to-one relationship between applicable entities while it was, in fact, in a one-to-many relationship. At some point, the database evolved, and this issue was corrected, resulting in an unfortunate situation where it became unclear where the actual data could be found. The morphing of the database structure is a witness testimony of, not only, the technology changes between 1999 and

2012, but also of how record-keeping changed during that period. The use of associated software with the record-keeping system showed changes from Lotus Word Pro to Microsoft Word (.doc) that was subsequently superseded by the OOXML format (.docx). Earlier, spreadsheet documents were rare but became a more common occurrence later. The record-keeping process went from being a combination of paper and digital to record-keeping becoming more and more digital, especially as the use of email became more prevalent. Before 2006 the system only stored outgoing documents electronically; incoming documents were registered within the system but remained paper-based. After 2006, incoming documents were scanned and stored within the system. As time went by, incoming documents went from having a paper form to being born digital. As with any system, it is expected that minor functional updates will see small changes to the underlying database, while major changes in functionality will result in substantial changes to the database; therefore, it is natural that the database structure morphs over time. Today, tools like liquibase[3] exist that nicely document database evolution. This type of useful documentation lacks for legacy systems.

It is commonly accepted that information should be treated as an asset for an organization (Dakova et al., 2018). Still, it can be difficult to see the requirement to understand the database asset at an Entity-Relationship model level fully. It is easy to fall into a practice where the technical side of things remains a vendor issue. It is observable that this problem is particularly prevalent when moving from paper-based record-keeping to electronic record keeping as it can result in a significant impact on the database structure. However, once all record-keeping is electronic, such structural issues will likely be less of a problem going forward.

A recommendation here is that IS managers should ensure they have updated documentation regarding the evolution of their databases. This should be done to provide insight into the gradual changes in technology that ultimately become a challenge for migration and preservation. Even a relatively short period of 13 years can see relatively significant changes to a database. The availability of up-to-date technical documentation should be part of the acquisition process. For older systems, the vendor may have ceased trading, been bought up or for some other reason, no longer supports the software. In the worst case, technical documentation may no longer be available, and the process of migration can become unnecessarily complicated.

Database issues

When analyzing the databases, it was discovered that the use of referential integrity was not enabled, nor was the use of primary keys in place. Referential integrity is a vital database mechanism that ensures the consistency of interrelated records within a database. An example of this is that referential integrity can prevent a user from registering a document within a database unless there is a connection to a case file and a case handler. Referential integrity can and should be used to prevent the occurrence of "orphan" records. The main issue with "orphan" records is that the Noark standard explicitly prohibits their existence, and when testing the output of a migration, the test will fail if "orphan" records are present.

Interestingly though, for important records (case files, document identifiers, etc.), the technology did manage to successfully enforce a level of referential integrity through software, rather than at the database level. There were many examples where referential integrity issues caused problems, but for the most, these were related to third-party integrations rather than issues with the vendor's software. The municipalities used some third-party systems that integrated directly to the record-keeping system database, inserting data without necessarily understanding the consequence on later migrations. In essence, these third-party integrations created "orphan" records that resulted in data deviating from the standard. This is likely due to the lack of availability of technical documentation. There was also an instance where a small collection of records had been disconnected from the official collection of records because a user had changed a field linking the records to the correct context. Had referential integrity been enabled in this case, then this would not have been possible. This is a problem because such records may not be identifiable and retrievable when later searching for data. It was also noted that, in a few instances, the lack of primary keys resulted not only in duplicated data but also records that had multiple meanings. The system allowed incoming documents to be identified as "*incoming*" with a status value "I". But the database had registered status "*I*" as being used for both "*incoming*" and "*incoming job application*", which have slightly different meanings. Duranti (2001, p. 272) notes that the lack of unique identifiers can raise issues regarding the perceived authenticity of records.

Given that referential integrity and the use of primary keys were not in place, it could be expected that questions regarding the authenticity of the records with these databases could be raised. Despite the lack of primary keys within the database, the software did a surprisingly good

job of ensuring the uniqueness of both case files and document registrations. These are the most important objects when considering a Noark-based preservation migration. As such, there is no real criticism that the technology was unable to ensure the authenticity of the records. The lack of referential integrity and the use of primary keys became more of a cost issue during the migration project as opposed to being an authenticity issue. Considerable effort was used to detect and remove duplicates as well as ensuring consistency between "orphan" records.

There were many minor issues with the database that mainly became a cost issue to rectify. A common issue was how the *null* value was often used to indicate a Boolean "*0*", but this was not implemented consequently. The Noark 4 standard often required a Boolean value to be set as either a "*0*" for false or a "*1*" for true, but intermittently *null* was used to indicate "*0*". However, in database theory, null means that no value has been set, and its use raised a type of philosophical question where we had to ask if it is acceptable for migration software to interpret some *null*-value fields. A particularly notable instance is where Noark uses a field, *publiclyExempt*, to define whether or not a record is publicly available for FoI requests as the law prohibits the publication of private or sensitive information. An example of such information can be found in an application from an older person applying to live in a municipal care-home where information about their ability to take care of themselves would be part of the application. When migrating records from the database, software interpretation of *null*-values as "*0*", can potentially result in a case file wrongly being published. The migration software is then, in essence, undertaking an evaluation that really should be undertaken by a person based on an inspection of the contents of the case file. While the issue may seem like a minor technical issue, it interrupted the migration project and became an annoying cost issue. Had the vendor simply used "*0*" instead of *null*, then there would not have been an interruption.

A recommendation for IS managers within this sphere is that it is essential to understand the conceptual model of the database and how it is implemented. In some cases, one simply cannot enforce referential integrity throughout the database, especially if the conceptual model requires a large degree of flexibility in the way records are interconnected. But it is disreputable to have neither the use of referential integrity nor primary keys in use in a modern software system built on top of a relational database. Responsibility for a technology includes understanding how that technology uses the database.

Deviation from standard

As noted, the technology claimed to comply with version 4 of the Noark standard. A few database tables did follow the naming convention outlined in the standard, while others had names that bore no similarity to the standard. Some of these strange table names are illustrated in Table 2.

Table 2

Examples of table names that are not in compliance with the Noark 4 standard

Noark 4	System name
JOURNPOST	DGJMJO
NOARKSAK	DGSMSA

Sometimes column names bore no similarity to the naming specified in the standard. Examples of this are shown in Table 3.

Table 3

Examples of column names that are not in compliance with the Noark 4 standard

Noark 4	System name
SA_TGKODE	UNTOFF
SA_UOFF	HJEMMEL
SA_TITTEL	INNH1

The examples depicted in Table 2 and Table 3 illustrate some of the mappings to get the database of the system to follow the Noark standard.

In some cases, the standard required a count field, e.g., the number of records associated with a case file. This field was missing in the database, but could quickly be produced with an additional SQL-query. There was one instance where two columns were swopped. This was not a significant issue but did require documentation. It was also noted that sometimes the database had differing data types and string field lengths than the standard expected. There were a few instances where the Noark 4 standard specified a column length to be VARCHAR(10), while the database implemented the column as VARCHAR(15). This was a potential problem when testing the output of the migration for compliance.

While these are small structural issues, they consume a lot of time when trying to undertake a migration. They raise uncertainty, increase documentation requirements, and slow

down the migration process. A recommendation here for the IS managers is to make sure that documentation showing how the database implements the applicable standard is readily available. Such documentation, however, can be difficult to procure as some vendors may see technical descriptions of the database as proprietary and confidential business information.

Incorrect system use

Two particular examples detailing how the technology had been used incorrectly were discovered. The first was concerning the use of encrypted files; the second was regarding document versions. During the migration process, it was discovered that some of the documents were encrypted. This was unfortunate as there was no documentation about the encryption algorithm or the key to decrypt them. The municipalities were not aware that the technology actively supported encryption. Of the 850 000 documents, there were just a few hundred documents that were encrypted. Still, it was a contentious issue as the technology should never have exposed the functionality to end-users. The original vendor had to decrypt the files for the municipalities. The second example was related to the technology's ability to store multiple versions of a given document, in essence, tracking how documents evolved. This became a problem as the project was informed that versioning functionality was not in place, but the database showed a few instances (<10) of multiple versions of documents. Such misunderstandings resulted in conflicting explanations and interpretations, causing delays to the migration process.

A recommendation here for an IS manager is to ensure that full knowledge of the technology and its capabilities are essential when undertaking a migration. A technology with hundreds of users will likely use the technology in a manner the IS manager is unaware of. Documentation and functionality testing are essential to combat this.

Missing data

There were various examples of missing data in the database. An example of this is how the Noark 4 standard supports some additional optional modules of record-keeping functionality that can be employed. In one instance, four weeks of data from an optional module was found. Upon investigation, it was discovered that the municipalities had tested some of the optional modules but decided not to use them after all. This information was not documented before the migration process. When a database contains records from a software module that subsequently is discarded and is limited to a testing period of four weeks, it quickly creates challenges for long

term preservation, in particular with regards to the perceived authenticity. If portions of the data are missing, how will future users (in 50 years) interpret the authenticity of the migrated data? It raises a question about what else may be missing, or what other intended or unintended changes happened during the migration process, and as such, future users might conclude that the migration process was likely to have been flawed. The authenticity of the entire migrated database may become questionable. Another example of missing data is where an administration table that contained information about users and group membership was sporadically missing required records. When a user was added to a group, the technology should have recorded the identification of the administrator that added the user to the group. This information was often missing. There was no reasonable explanation put forward as to why this information was missing, but it was believed likely to be a result of a software bug that intermittently came and went away. A single issue like this is likely to be seen as a trivial issue, but when many minor issues are prevalent in the database, the authenticity will become questionable.

A recommendation here for the IS managers is to make sure that the procurement of a technology includes a requirement of a test-suite with a minimum set of tests to ensure compliance with the applicable standard. It is also worth noting that unless there is a formal requirement that a technology has to adhere to a given data model for migration or preservation, it may be challenging to discover potential problems until the migration is underway.

Automated Document Migration

Document migration, from production to preservation formats, is an issue most IS managers will have to deal with at some stage. This is particularly true when documents adhere to older file formats, e.g., Lotus Word Pro or WordPerfect. There is a general requirement that preserved documents are migrated to a format that is suitable for long term preservation. PDF/A (ISO 19005-1:2004, 2004) is a good example of a file format that is preservation friendly as the document and its contents are self-contained, ensuring that the contents can be rendered in the future. A document migration process can be risky, as it is often automated, and the outcome of the process can be difficult to verify. An analysis of the various file formats within the document collection is detailed in Table 1. At a glance, such a list may not seem daunting at all, but each additional file type comes with an organizational knowledge requirement that can be expensive to develop and maintain. As pointed out by Suri and El-Saad (2017, p. 2), the competencies required to handle various file formats are extensive; the available software to undertake the

conversion is limited, and with batch processing, the result is often unverifiable. File format heterogeneity combined with volume, cost, time, and technological obsolescence quickly become issues that impede automated migration. File formats that have no obvious preservation equivalent, e.g., DXF and DWF (a CAD format for storing two- and three-dimensional design data and metadata) and ZIP files (contents have to be extracted and checked) quickly become a hinder to automation and batch processing. Their existence in a system can add weeks or months of work onto an automated migration/preservation process due to various inquiries and investigations to be undertaken and ultimately cause delays and significantly increased costs. Ironically, small numbers of non-standard file types can be quite difficult and expensive to deal with, and their migration cost can be as high as large numbers of more widely adopted file formats. An example of this is dealing with the DWG files as opposed to the doc files. The doc files were batch-converted using known migration software, while the DWG files were migrated manually.

Volume and time are related, as the volume of documents naturally increases over time. When the system went into production around 1999, Lotus Word Pro (LWP) was the predominant word processor in use. Over time the municipalities changed to Microsoft Office, and the Word-format became the predominant file format stored in the system. As time went by, the volume increased, but slowly the software to handle LWP files fell away. As such, technological obsolescence becomes an issue that causes unforeseen difficulties in the future. These issues quickly become a hinder when attempting automated migration processes.

A concrete example of this was observed when attempting to convert LWP to PDF/A and how neither MS Office, OpenOffice nor LibreOffice software packages were successfully able to convert all the LWP files. OpenOffice was able to open and convert some LWP files but often crashed, rendering the conversion attempt futile. In some cases, the conversion attempt resulted in a corrupted PDF document, and it was noted that in one instance, the conversion of a single-page LWP document resulted in a 93-page PDF/A document full of binary symbols. OpenOffice did not report any issues or errors when attempting this conversion. In the end, it was decided not to convert all the LWP documents, as an LWP file is a combination of plain text and binary symbols. Hence, a lot of the written content of the document is retrievable using manual labor with a text editor. It was left to future users to extract the document content manually if required. A recommendation from this work is that an IS manager in charge of record-keeping, or other

digital libraries should have an ongoing strategy that identifies heterogeneity and volume and should be able to identify potential hinders for migration. It is the slow changes over time that causes most of the problems.

Discussion

Deviation from the standard was the single largest source of problems for the migration project, and a lot of the issues can be traced back to this issue. This is an important issue as each small deviation may seem like a minor problem. Still, many minor issues become a big problem and quickly take up a disproportionate amount of time when migrating records. The main reason for this problem is due to the lack of a compliance testing regime. Standards compliance for Noark has traditionally been a self-verification process where the vendor simply informs the standardizing body that compliance is in place. The standard itself opens up for deviations and states that "the data model is intended as an example for those who need to develop Noark-based systems, and is not a requirement in itself" (Noark 4, 1999, p. 9). The standard follows up and notes that it should be possible to produce a valid migration, and compliance with the data model as such is obvious. When a standard fails to enforce actual standardization, the kind of problems we detail here are inevitable. In hindsight, it is easy to criticize that the standard did not enforce the use of the data model consistently, but there was likely a necessity to ensure backward compatibility with earlier versions of Noark (versions 1 through 3).

Migration, however, is a controversial process, and "a migration will always result in some losses" (Wheatley, 2001, p. 2). As such, it is crucial to understand the implications of every migration action fully. Any errors during this phase can have dangerous consequences for preservation and the ability to locate information in the future. It can be difficult for an IS manager to focus on an inevitable migration process when a migration is not likely to be an issue until 13 years into the future. Technology can change considerably over 13 years.

A recommendation here is that the IS manager should approach software acquisition from a life-cycle perspective where end-of-life or general migration issues are part of the acquisition process.

The next important factor when looking at the migration process is how volume, heterogeneity, and time cause problems when dealing with document collections. Volume quickly becomes a problem as it precludes manual processing, and essentially only automated approaches are applicable due to costs and for efficiency reasons. However, the results of

automated processes are not always verifiable, and this can result in both corruption and loss of data. Heterogeneity compounds this problem as the greater the variation in file formats, the greater the associated costs when undertaking a migration. Volume increases naturally over time, but time also sees technology changes. As such, these interrelated factors are a significant complex combination that affects a preservation migration.

File format heterogeneity is a dilemma for the IS manager. Capturing all document types will give a richer and more complete set of records, but with increased costs for migration. Limiting the allowed file types will preclude documents coming under record keeping control, but make migration easier. The former is likely to be the correct approach, even though the final migration costs are unknown. The recommendation to the IS manager is to take responsibility. A vigilant IS manager will, from the outset, know that this is an underlying potential problem that must be kept in check and will have procedures in place to identify issues related to the creation of preservation versions of documents.

The third impacting factor hindering migration is database evolution, but it is the lack of updated documentation that is a problem. A changing domain model where data has multiple locations during the lifetime of the database means data can quickly be lost during migration. The problem can be mitigated with a testing regime, documentation, and a requirement that the software has built-in migration extraction capabilities that can be run at any time.

The remaining issues, e.g., database issues, incorrect system use, missing data, detailed in this paper did have consequences but were, in a way minor, compared to the three detailed above. As the database was not in compliance with the standard, migration issues emerged. From an archival perspective, the migration is *valid* but potentially *incorrect* as the contents of the database were extensively processed to create a migration. The problem here is that future users will see such a migration and believe that the technology was in compliance with the standard. In essence, the migration is counterfeit, a lie about the technology. As such, there is also a need to preserve the original database that is seen as *invalid* (according to the standard), but *correct*. Technically, this would tell the truth about the technology, but would also leave problems and missing data. As there is a potential conflict between the first and second approaches, there is a requirement to store a copy of the original database. From a preservation and authenticity perspective preserving variations of the same data may seem redundant and overkill. However, this should be seen as the archival cost of deviating from the standard.

Conclusion

The results of an empirical study of issues related to migration for preservation purposes are presented along with a discussion of their consequences and recommendations for IS managers working with such data. Hence, this paper presents a rare insight into the migration of municipality record-keeping databases. Even though the project is limited to a Norwegian standard, the results are relevant to the broader IS community where migration for preservation purposes is to be undertaken. Six Norwegian municipality databases were analyzed, and a migration was undertaken to deposit the records and associated documents with an archival institution. The project was mostly successful in undertaking an actual migration but was unable to provide a definitive answer on whether or not a complete migration was enacted. The migration project achieved coverage of about 95% and concluded that the potential work input to resolve the remaining 5% was not worth the required effort. Technically, the extraction was inadequate according to national guidelines, but the extraction was accepted with the documented limitations. In terms of the ISO 15489:2016 (2016, p. 9) migration definition, that migration is an "act of moving records from one system to another while maintaining the records' authenticity, integrity, reliability and usability", one can argue about the authenticity of a migration where there are so many problems.

On the one hand, the migration project managed to increase authenticity as the original database is now extremely well documented for future users, and there exists a copy of the records in a structure that is mostly compliant with the governing standard. On the other hand, when considerable processing has to be undertaken, one can be left in doubt about the integrity and quality of the records. Any future authenticity evaluation of the extracted records will likely be based on the amount of documentation that is available. The migration is deemed by all parties to have been successful, and the material is considered authentic but tainted by the identified problems.

The results of this study are not transferable to Noark systems by other vendors, as little is known about their implementations of the standard. Instead, they point to an issue with legacy systems that have a longer usage period in production, that see an intermittent morphing of the underlying database concerning software updates. It was observed during this migration project that the IS managers have neither the time nor the technical ability to concern themselves with the underlying database. They focus on the daily record keeping task. In Norway, the record-

keeping profession is perhaps putting too much reliance on the vendors of record-keeping systems and not taking enough responsibility.

The most important consequence of this work is a recommendation that the IS manager takes greater responsibility in understanding issues related to migration when acquiring a particular record-keeping system. If the system claims compliance with, e.g., Moreq2010, it must be possible to verify compliance. For most systems, a migration will occur at some time, either when replacing the software with another implementation or when the system is no longer in use, and the records will be subjected to a preservation process. The longer one waits, the more expensive the process is likely to be.

References

Becker, C., Antunes, G., Barateiro, J., Vieira, R., & Borbinha, J. (2011). Modeling digital preservation capabilities in enterprise architecture. In *Proceedings of the 12th annual international digital government research conference: Digital government innovation in challenging times* (pp. 84–93). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2037556.2037570 doi: 10.1145/2037556.2037570

Bisbal, J., Lawless, D., Wu, B., & Grimson, J. (1999). Legacy information systems: Issues and directions. *IEEE Software, 16* , 103-111.

Brodie, M. L., & Stonebraker, M. (1995). *Legacy information systems migration: Gateways, interfaces, and the incremental approach*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Dakova, J., Antunes, P., & Chiu, Y.-T. (2018, 01). A pluralistic approach to information valuation. In *PACIS 2018 Proceedings*.

Duranti, L. (2001). Concepts, principles, and methods for the management of electronic records. *The Information Society, 17* (4), 271-279. Retrieved from https://doi.org/10.1080/019722401753330869 doi: 10.1080/019722401753330869

Duranti, L. (2005, 01). The long-term preservation of accurate and authentic digital data: The interpares project. *Data Science Journal, 4* . doi: 10.2481/dsj.4.106

Hagen Sataslåtten, O. (2017, nov.). The norwegian noark model requirements for edrms in the context of open government and access to governmental information. *Tidsskriftet Arkiv, 8* (2). Retrieved from https://journals.hioa.no/index.php/arkiv/article/view/2485 doi: 10.7577/ta.2485

International Organization for Standardization. (2004, March). *Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A)* (Vol. 2014; Standard). Geneva, CH: International Organization for Standardization.

International Organization for Standardization. (2010, March). *Information and documentation – Principles and functional requirements for records in electronic office environments – Part 1: Overview and statement of principles* (Vol. 2010; Standard). Geneva, CH: International Organization for Standardization.

International Organization for Standardization. (2010, March). *Information and documentation – Principles and functional requirements for records in electronic office environments – Part 3: Guidelines and functional requirements for records in business systems* (Vol. 201o; Standard). Geneva, CH: International Organization for Standardization.

International Organization for Standardization. (2011, March). *Information and documentation – Principles and functional requirements for records in electronic office environments – Part 2: Guidelines and functional requirements for digital records management systems* (Vol. 2011; Standard). Geneva, CH:

International Organization for Standardization. International Organization for Standardization. (2012, March). *Information and documentation – Digital records conversion and migration process* (Vol. 2012; Standard). Geneva, CH: International Organization for Standardization.

International Organization for Standardization. (2012, March). *Space data and information transfer systems – Open archival information system (OAIS) – Reference model* (Vol. 2012; Standard). Geneva, CH: International Organization for Standardization.

International Organization for Standardization. (2016, March). *Information and documentation – Records management – Part 1: Concepts and principles* (Vol. 2016; Standard). Geneva, CH: International Organization for Standardization.

Lin, L. S., Ramaiah, C. K., & Wal, P. K. (2003). Problems in the preservation of electronic records. *Library Review, 52* (3), 117-125. Retrieved from https://doi.org/10.1108/00242530310465924 doi: 10.1108/0024253031046592

Lorie, R. A. (2001). Long term preservation of digital information. *In Proceedings of the 1st acm/ieee-cs joint conference on digital libraries* (pp. 346–352). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/379437.379726 doi: 10.1145/379437.379726

Lübeck, M., Geppert, D., Nienartowicz, K., Nowak, M., & Valassi, A. (2003, 05). An overview of a large-scale data migration. In *Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSS'03)*. doi: 10.1109/MASS.2003.1194835

Maatuk, A., Ali, A., & Rossiter, N. (2008). Relational database migration: A perspective. In S. S. Bhowmick, J. Küng, & R. Wagner (Eds.), *Database and expert systems applications* (pp. 676–683). Berlin, Heidelberg: Springer Berlin Heidelberg.

Martens, A., Book, M., & Gruhn, V. (2018). A data decomposition method for stepwise migration of complex legacy data. In *Proceedings of the 40th international conference on software engineering: Software engineering in practice* (pp. 33–42). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3183519.3183520 doi: 10.1145/3183519.3183520

McKemmish, S. (2001, Dec 01). Placing records continuum theory and practice. *Archival Science*, 1 (4), 333–359. Retrieved from https://doi.org/10.1007/BF02438901 doi: 10.1007/BF02438901

Mellor, P., Wheatley, P., & Sergeant, D. (2002). Migration on request, a practical technique for preservation. In M. Agosti & C. Thanos (Eds.), *Research and advanced technology for digital libraries* (pp. 516–526). Berlin, Heidelberg: Springer Berlin Heidelberg.

Motsaathebe, L., & Mnjama, N. (2009). Managing court records : a survey of records-keeping practices in selected countries [Journal Article]. *Mousaion, 27* (2), 132-153. Retrieved from https://journals.co.za/content/mousaion/27/2/EJC78967

Rahman, A. U., David, G., & Ribeiro, C. (2012). Siard archive browser. In P. Zaphiris, G. Buchanan, E. Rasmussen, & F. Loizides (Eds.), *Theory and practice of digital libraries* (pp. 496–499). Berlin, Heidelberg: Springer Berlin Heidelberg

Ross, S. (2012). Digital preservation, archival science and methodological foundations for digital libraries. *New Review of Information Networking, 17* (1), 43-68. Retrieved from https://doi.org/10.1080/13614576.2012.679446 doi: 10.1080/13614576.2012.679446

Sirevåg, T. (2014, nov.). Utviklingen av noark-standarden 1984 – 2008. *Tidsskriftet Arkiv, 5* . Retrieved from https://journals.hioa.no/index.php/arkiv/article/view/1149 doi: 10.7577/ta.1149

Suri, R. E., & El-Saad, M. (2017). Lost in migration: document quality for batch conversion to pdf/a. *Library Hi Tech, 0* (0),. Retrieved from https://doi.org/10.1108/LHT-10-2017-0220 doi: 10.1108/LHT-10-2017-0220

The National Archives of Norway. (1999, January). *NOARK-4 Norsk arkivsystem Versjon 4* (Vol. 1999; Standard). Oslo NO: The National Archives of Norway.

Wheatley, P. (2001). Migration: a camileon discussion paper. *Ariadne, 29* (2).

Footnotes

[1]https://www.arkivverket.no/forvaltning-og-utvikling/noark-standarden

[2]https://lovdata.no/forskrift/2017-12-19-2286/§3-1

[3]https://www.liquibase.org/