N4OK: Kvalitetsvurdering av avlevert materiale

1. Overview

This document is the result of a year long project on data quality and archive that involved researchers and practitioners from the fields of data quality and archives. HiOA led the project and undertook 4MM of work. KDRS and IKA Kongsberg were partners contributing 1MM of work each and Dr. Markus Helfert from DCU was subcontracted in as DQ expert to guide the project. The project has analysed what Data Quality is from a Noark 4 records management and archives perspective. This report in divided up into the following Chapters:

Chapter 1: Overview
Chapter 2: DQ tools (Deliverable 1)
Chapter 3: DQ Dimensions (Deliverable 2)
Chapter 4: DQ Run (Deliverable 3)
Chapter 5: DQ Comparison (Deliverable 4)
Chapter 6: DQ Workshop
Chapter 7: Findings
Chapter 8: Cost report

We present here an overview of what the reader can expect from reading this report:

Chapter	Description	Result
2	DQ Tools	A definition on data quality is presented along and which tools can be used to measure DQ that follows the definition. A concise descriptions of relevant tools is presented.
3	DQ Dimensions	A number of relevant data quality dimensions and how these dimensions relate to Noark 4 are explored. This work goes deeper and looks at data quality from a number of different roles or users that will to some degree experience data quality as an issue.
4	DQ Run	What DQ is within a Noark 4 context is explored and documented. A pragmatic approach to measuring DQ from a Noark 4 extraction is explored
5	DQ Comparison	Is there are difference in DQ between the database and the extraction is explored and

6	DQ Workshop	It was decided that the project should reach out to the community that hopefully will benefit from this project and we held workshop in conjunction with a KDRS meeting in Trondheim in June 2013. Additional input resulting from the workshop is presented and included into the project. An evaluation of the project results is presented. The community report that the project results are relevant for them.
7	Findings	The project is first summarised in a concise manner before a presentation of additional findings is presented along with an approach to future work.
8	Costs	Cost report detailing project financial details as the are today.

Four project meetings were held, led by each partner in turn during the project. Regular meetings and discussion were often held using Skype. Google docs was the chosen collaboration platform for working on documents. All reports and collaboration was undertaken in english.

The findings presented in this report are based on observations and discussions and do not necessarily reflect the official opinions of the organisations that partook in the project. No conclusion should be drawn about the compliance to the Noark standard or the quality of any Noark 4 system. The project partners would like to thank Kulturrådet for supporting this project.

2. Report on tools (Deliverable 1)

Introduction:

This chapter documents various tools available that could be used to measure data quality for Noark 4 extractions. We do not aim to cover every single tool, instead we look at the general types of tools available. The report is based on a qualitative case study where we have investigated which available tools can be used to measure data quality. An additional non-binding constraint is that the tools should be either open source or at least free to use for a limited period of time. An IKA will not have a significant budget to spend on measuring data quality so expensive proprietary tools will not be of much use. As an example, proprietary software vendors like Microsoft and Oracle have often per-CPU or per-core licensing of their database products. These products can be used to measure some aspects of data quality but the cost can prove too expensive for smaller archival institutions. We first start by setting the context for data quality and discussing what it is.

What is Data Quality

Data Quality is a quantification by some means that assess the quality of data in a system (Wang et al. 1998; Wand and Wang 1996; Pipino et al. 2002). By quality we mean a property of the data that identifies a degree of excellence. When it comes to paper records, it is very easy to assess the quality of the underlying paper. It is however far from trivial to identify the quality of the written contents of that paper and a large number of paper based records poses severe limitations in terms of processing time. But is it any easier to identify the quality of electronic records? Techniques to measure data quality can handle large amounts of electronic records and process them and measure their data quality (Knight and Burn, 2005), something we could not achieve with paper based records.

An electronic record can loosely be understood as information captured in electronic form, but when it comes to records management and archives, a record has a more defined meaning. Provenance, context, structure and fixity are central elements of what defines an electronic record and a record often documents that some kind of transaction has taken place. Electronic records are, in most cases, stored in some sort of relational database and it is easy to see how records management and databases fit well together. The classic archival structure: fonds, series, file, record and the defining elements of an electronic record: provenance, context, structure and fixity can easily be implemented in a relational database. Relational databases do not by themselves ensure good quality of the data within. The existence of what we would call *bad data* could for example be a result of bad system design or improper data entry. What is undeniable is that data quality is an issue that is evident in the data stored within the tables of a database. Figure 2.1 illustrates how the data quality issue manifests itself in a database and how the source of the problem can be anything from system design/modelling or incorrect data processing.

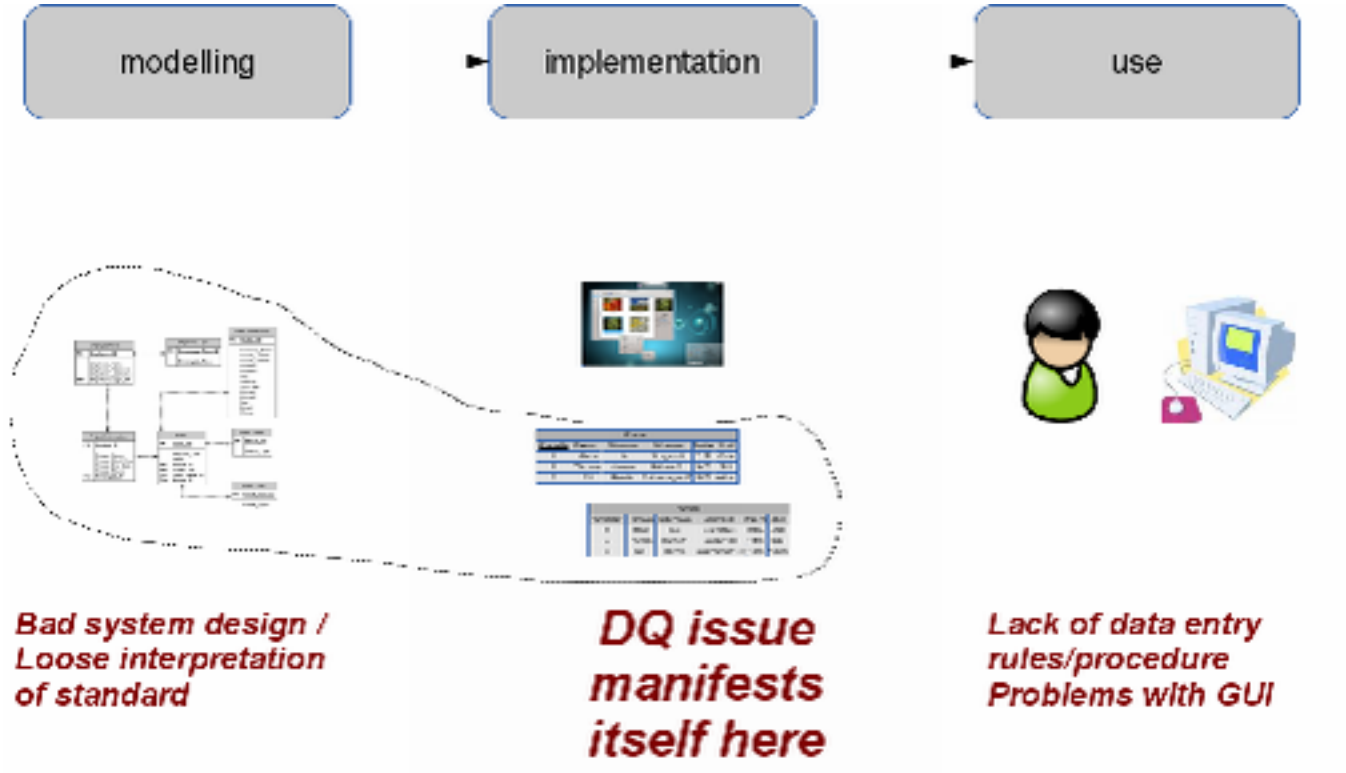

Figure 2.1: *Illustrating data quality as in issue in the tables of a database*

There can be many reasons for the manifestation of bad data in a system but two primary sources are often attributed to substandard system design and incorrect data entry (Pham and Helfert 2007). Substandard system design can be a result of a modelling process that fails to capture the real world scenario correctly but can also be a result of an incorrect interpretation of a model or the standard on hand. Problems with data entry could for example be a case handler entering incorrect data or due to a flaw in the user interface. There is often no single cause, but a myriad of causes which result in bad data quality.

The classic software development cycle starts with a data modelling process. First a real world scenario is modelled and that model leads to the development of a software system. An example how an inadequate data modelling job could lead to bad data is evident in what is known as the year 2000 problem. For many years two digits was more than enough to represent a year according with the real world scenario. Designers never foresaw that the software would still be in use around the year 2000 and that the change from the year 99 to the year 00 could suddenly cause problems. An example of this is dealing with age. Up until the year (19)99 age could be calculated correctly as todays year minus year of birth. So in (19)99 the calculated age of a person born in (19)64 could be 99-64 or 35 years. In the year (20)00 the same person would now be -35 years 00-35. In this example an inadequate understanding of the data requirements and its use leads to bad data quality.

In a perfect data quality world the systems would model the requirements perfectly, case handlers would have ample time when handling their cases and ensure data is correct and consistent. Alas we do not live in a perfect data quality world (Wand and Wang 1996). Software providers update their software because of bugs and changing user requirements. Within a system, users come and go and case handlers have a limited amount of time to undertake their job. Data quality is often not an issue case handlers are concerned with and the long term perspective of the data is also far from their minds.

For the purpose of this project we are concerned with Norwegian municipality data and limit ourselves to data within a Noark 4 context. When looking at the life-cycle of data in a Noark system we can identify three distinct phases. These phases are shown in Figure 2.2

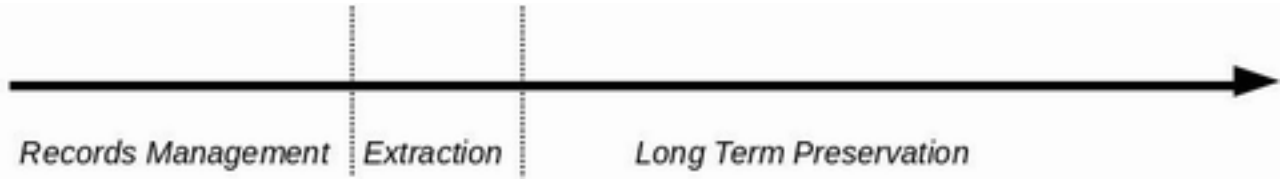

Figure 2.2: *The three traditional phases data in a Noark system experiences*

During the records management stage records are often born digital and a municipality RM database will experience a continual growth of the number of cases over time. Over time the number of cases can be very large and the ability to correct records with bad data quality is severely limited due to volume. When it comes to the extraction phase, the municipality deposits a copy of its data in a given format with an archival institution. The volume here is so large that it can be difficult to correct any data quality issues (see also for instance Data Quality in Helfert 2002; Cappiello 2003; Ballou and Tayi 1999). When the data enters its long term preservation stage the disconnect between the creator and the maintainer is so large that it may be impossible for the archive institution to fix bad data. The archive institution will find itself in a difficult situation. They will maintain and provide access to this data for the long term and if the data has bad quality it may result in significant additional costs to ensure access to the data through a process of manual processing to answer questions from the public about the data.

For the purpose of this project we define data quality as "***the degree to which data in a system reflects the real world scenario the data represents and is usable***". The Norwegian equivalent of this definition is "***i hvilken grad data i et system er i overenstemelse med det virklige scenarioet dataen representerer og er brukbar***".

This definition is not as fluid as other definitions on data quality (e.g. Wang and Strong, 1996; Wang et. al. 1998; Scannapieca and Batini, 2006) but is an attempt to correctly define data quality given the nature of the type of data. This data is very domain specific, public administration data that is realised in a Noark 4 system. The data experiences three distinct

phases in its lifecycle and should exist for forever. The modelling of the real world scenario is given to the software developer in the form of a standard so compliance with the standard should, *in theory*, result in a negligible effect of negative data quality (Wand and Wang 1996). The concept of usability can be difficult to gauge, as it is not necessarily easy to identify who the user of the data will be. During the records management phase the user is clear and explicit. Employees of the municipality are the defined users. When it comes to long term preservation, only the archive institution can look at the data for the first 60 years and after that, in many cases, the data is available to the public. It is difficult to define who the users of the data will be in 500 years and what requirements these users will have.

If we attempt to model or capture the concept of a user then it is easier to define what we mean by “fitness of use” or usable. The definition of a user can be simplified and defined indirectly in terms of the standard that the data complies with. We can roughly assume the requirements users in the future will have by defining what we capture during the records management phase and deposit at the archival institution. Two types of users are already clear to us. The first needs to process and link data for research purposes while the second user will want access to individual records. Identifying trends from the statistics is a good example of how we expect people in the future to research Noark records. Data Quality then will then be important issue across collections as the data has to be comparable. When it comes to locating and viewing individual records two aspects will be important, the ability to locate records and the the ability to understand them. Traditional search algorithms have come far enough to handle this but traditional search algorithms do not perform well in cases where data quality is poor. The ability to understand the records is also extremely important and poor data quality may hamper the ability to understand individual records. Good data quality is needed by both parties.

We measure data quality in what we call data quality dimensions (Wang and Strong 1996; Pipino et al. 2002). A data quality dimension is particular measurement of quality and these dimension are often categorised differently (Ge and Helfert 2007; Miller 1996). There are a number of different categorisations in use, but we will loosely define data quality in terms of the following three categories:

- Objective Data Quality
- Subjective Data Quality
- Process Data Quality

This categorisation is not the only method of categorising data quality dimensions but at this stage in the project it serves to create a foundation to guide further work. Subjective data quality is often assessed using a survey where the users of the data are often asked how they perceive the quality of the data they use. While this approach guages the fitness-for-use aspect of the data, it is also very subjective to the person answering the survey (Fehrenbacher and Helfert 2012), . A case handler in a municipality with 15 years of experience may easily perceive the data on the screen in front of them as being of high quality because they have a lot of background knowledge and experience. A recently employed case-handler may look at the same screen of information and validly perceive the same data as having low quality. This

anomaly can mean it is harder to interpret the value of subjective data quality but its strength is that it puts the user of the data centre stage when assessing data quality. It is hard to see the benefit of subjective data quality when considering the long term perspective on Noark 4 records. The user that today assesses the quality of Noark data as being of high quality, will not be present in the future, in for example 10, 100 or in 1000 years. Also the future user might perceive some Noark 4 data that was previously judged as having high data quality as now having low data quality. As the user is difficult to identify we need to focus objective data quality assessments.

An objective data quality analysis is one that is independent of users and when run on the same data will always give the same answer. Objective data quality assessments can follow the data through its lifetime and indirectly is part of defining the requirements of the user. This approach is very much in keeping with the thinking behind the OAIS model where an archive institution is expected to have identified a designated community for their holdings. As such data quality in an archive context can be defined in terms of the designated community and it therefore becomes important to be in a position to to objectively measure the quality on the basis of the expectations of the designated The data quality measurements should as far as possible be objective in nature as they can readily and independently be undertaken over time. The amount of data can be and will be relatively large. Manual processing of the data is not an option and a focus on data quality can quickly be an investment that pays itself of by avoiding costly processing in the future.

This project deals with data that was instantiated and existed according to the Noark 4 standard. During the project we may discover problems with the standard or uncover misinterpretation of the Noark standard in the system being evaluated. We have no desire to to focus on whether or not the Noark 4 standard is in compliance with the real world scenario it is representing.

Tools to measure Data/Information Quality

When it comes to measuring data quality we have already defined the context as data within a database. From a Noark 4 point of view, we are really only concerned with relational databases. So the tools we can use to measure data quality will have some ability to process data in a relational database. From our analysis we have seen there a number of different categories of data quality tools that can be used to measure data quality of Noark 4 data. We categorise them as:

- Internal Database data quality tools
- External Business Intelligence data quality tools
- External standalone data quality tools
- Domain specific tools

A number of DBMS systems have internal data quality tools that can be used to assess data quality. MSSQL Server and Oracle are good examples of this. External tools to measure data quality include DQ Tools that are often found as part of Business Intelligence software or ETL[1] software. There are also stand alone data quality tools. Domain specific data quality tools are developed for a specific domain and we found two Noark 4 tools that have data quality functionality.

Internal Database data quality tools

Both Microsoft and Oracle have data quality tools as part of their database software. MSSQL Server is a popular database developed and sold by Microsoft. MSSQL Server includes a data quality service as part of its database offerings. These tool provide the following types of functionality: data cleansing, semantic duplicate matching, data quality verification with datasets and data profiling. Both MSSQL Server and Oracle are extensively in use by parts of Norwegian public administration.

External Business Intelligence data quality tools

There are a variety of business intelligence tools that have some form of data quality analysis in them. Business Intelligence tools are often used by management when it comes to trend analysis and when they need to make long term decisions. Good data quality is the foundation for business intelligence as it is difficult to make managerial decisions if the underlying data is bad. While business intelligence is interesting from an archive perspective we see it as a significant investment for Norwegian municipal archives to invest in this kind of software.

External standalone data quality tools

There exists a number of standalone data quality tools that work on any database. These are generic data quality tools and not domain specific, however they can be used with the Noark4

DQ 360 Data Cleansing Software Suite

This is a proprietary tool that includes functionality like profiling, validating. Parts of this tool could be used for the Noark 4 domain. Pricing information is not publicly available.

talend Open Studio for Data Quality

The software is a open source tool released under both the GPL and Apache freeprog licenses. The tool is built on top of the eclipse framework and can be used to carry out a number of data quality functions.

Domain specific tools

arkn4

arkn4 is a Noark 4 testing tool that can be used to validate a Noark 4 extraction. It was developed by Redpill-Linpro for the Norwegian National Archive and released under a GPL

[1] ETL stands for Extraction, Transformation and Loading

license. This tool has functionality that makes it a DQ tool. In particular it checks case statuses and referential integrity constraints.

URD

URD (Universal Relational Database) is a tool developed by Frode Kirkholt that allows an archive institution to import and maintain any database structure. This tool has functionality that makes it a data quality tool. It appears to have a greater array of data quality functionality than arkn4 including for example a check on the length of titles that case handler enter on registrations.

POSQL

Plain **O**ld **SQL** can be a very powerful tool when measuring DQ. While not SQL statements themselves are not domain specific, a set of SQL statements to measure data quality can be grouped. It is also very flexible in terms of reusability.

Tools Comparison

Tools built on top of a particular database (MSSQL Server or Oracle) are very useful for assessing the quality of data stored in those databases. The biggest drawback of this is that data quality functionality is generic in nature. The results from this project could probably be implemented in these databases, however we believe it will be better to make the results available for wider usage. There are plenty of business intelligence tools available that could be used to help detect data quality issues with Noark 4. These tools tend to be expensive and overkill from the Noark 4 perspective. They are not domain specific and their generic nature might mean it is difficult to get them to sufficiently work for the Noark 4 domain. The generic data quality tools have the pure data quality focus but also might be limited in what can be achieved as they might be too generic. POSQL is the simplest and most portable way to measure data quality. arkn4 and URD are tools that are domain specific that include a limited set of data quality functionality. Their focus is more to assess compliance with the Noark 4 standard than assessing data quality.
A second way to look at these tools is where they will be most useful. If data quality is a problem in Noark 4 during the records management phase then the non domain specific tools could be used to identify and correct problems. When it comes to extractions in an archive institution we will probably be most interested in domain specific tools. It will be overly cumbersome for an archive institution to have to constantly update the Noark 4 domain in a generic tool when the software provider inevitably upgrades the tool. We can also question if the non domain specific tools will be available in the future and as such domain specific are preferred.

Conclusion

The tools we have seen so far can easily be divided into two categories, generic and domain specific. Generic tools can be used on any domain, while domain specific tools are optimised for a particular domain. Most generic DQ tools can be adapted to measure for example parts of

the Noark 4 domain but it makes more sense for the project to work towards a domain specific tool. While the database DQ tools are impressive, implementing the results on a given DBMS will, to a certain degree, tie the future use of the results to that DBMS and may slow down any uptake of the results. However it is worth nothing that archivists and record managers should be aware of the functionality these systems provide as they might be able to use them to capture and correct DQ issues during the records management phase of the life cycle.
There are plenty of options that allow for an abstraction away from any given DBMS. BI and ETL tools can often connect to a variety of DBMS. However we argue against the use of BI or ETL tools unless they are already in use for other purposes. These tools tend to be expensive and generic in nature. Talend is an interesting open source tool that can be used to measure DQ for Noark 4. It is generic and offers a lot. However given the existence of arkn4 and URD we believe we should focus on domain specific tools. Both of these tools are open source and the results could easily be integrated to both.
If we briefly look away from Noark 4 to specialised systems DQ tools. The General Auditors report from 2010 states that it is within these systems that we anticipate we will see the brunt of problems in the future. A short term solution for these systems could be to raise awareness around data quality and develop strategies to measure and correct DQ issues over time. Record Managers could use existing DQ services in database software or tools like talend to measure DQ.
Our conclusion is for the project to continue and develop results in plain old SQL. This allows the results to be implemented in both URD and arkn4 but also portable to other scenarios. Many archival institutions do not have large budgets so to avoid any type of vendor lock-in, either to a software company or DBMS we recommend that DQ for Noark 4 should, as far as possible, be measured with plain old SQL statements.

References:

1. Ballou, D.P. and G.K. Tayi (1999) “Enhancing data quality in data warehouse environments”, Communications of the ACM (42)1, pp. 73-78.
2. Cappiello, C., C. Francalanci and B. Pernici (2003) “Time-related factors of data quality in multichannel ISs”, Journal of Management Information Systems (20)2, pp. 71-92.
3. Fehrenbacher, Dennis and Helfert, Markus (2012) "Contextual Factors Influencing Perceived Importance and Trade-offs of Information Quality," Communications of the Association for Information Systems: Vol. 30, Article 8.
4. Ge, M. and M. Helfert (2007a) “A Review of Information Quality Research” in Proceedings of the 12th International Conference on Information Quality 2007, Cambridge: MIT, pp. 76---91.
5. Helfert, Markus: Proaktives Datenqualitätsmanagement in Data-Warehouse-Systemen - Qualitätsplanung und Qualitätslenkung-, Logos-Verlag, Berlin 2002 (ISBN: 3-89722-930-7).
6. Knight, S. and J. Burn (2005) “Developing a Framework for Assessing Information Quality on the World Wide Web”, Informing Science (8)1, pp. 159-172.
7. Lee, Y., D. Strong, B. Kahn and R.Y. Wang (2002) “AIMQ: A Methodology for Information Quality Assessment”, Information & Management (40)2, pp. 133-146.

8. Miller, H. (1996) “The multiple dimensions of information quality”, Information Systems Management Spring (13)2, pp. 79-82.
9. Pham Thi, Thanh Thoa; Helfert, Markus:. 2007. Modelling Information Manufacturing Systems. International Journal of Information Quality, Vol. 1, 1, pp5-21.
10. Pipino, L., Y.W. Lee and R.Y. Wang (2002) “Data Quality Assessment”, Communications of the ACM (45)4, pp. 211-218.
11. Scannapieca, M. and C. Batini (2006) Data quality: concepts, methodologies and techniques, Berlin: Springer.
12. Wand Y. and R.Y. Wang (1996) “Anchoring data quality dimensions in ontological foundations”, Communications of the ACM (39)11, pp. 86-95.
13. Wang, R.Y. and D.M. Strong (1996) “Beyond accuracy: What Data Quality Means to Data Consumers”, Journal of Management Information System (12)4, pp. 5-34.
14. Wang, R.Y., Y. Lee, L.L. Pipino and D.M. Strong (1998) “Manage Your Information as a Product”, MIT Sloan Management Review (39)4, pp. 95-105.

Deliverable 2 : DQ Dimensions for Noark 4

Introduction:

This document reports on the DQ dimensions the projects believes are applicable to Noark 4 extractions when an archive institution is in the process of receiving an extraction.

Methodology and Approach:

The report is based on a qualitative case study where we have studied general data quality dimensions and how they can apply to Noark 4 data.

Setting the stage

Noark 4 is the fourth version of a records management standard that Norwegian public sector is obliged to use for records management. The requirement to use Noark is explicit under §2-9 of the regulations relating to public archives[1]. Noark 4 defines a records management system for public sector case handling and is very detailed, including a definition of relational database tables and attributes. The standard came in 1999 and was eventually overtaken by Noark 5 in 2008. There are many Noark 4 systems in use in Norway today.

To make this report easier to understand we define a number of common data products and user roles that exist within a Noark 4 context. This is done to avoid confusion of terminology.

1. A Noark 4 Case file
A Noark 4 case file (no:saksmappe) is an aggregation of information about a particular case and only information related to a case is stored within a casefile. Examples of metadata associated with a case file include *title*, *date created*, and *classification code*. A case file will include a number of data products in the form of registrations that link to documents.

2. A Noark 4 Registration
A Noark 4 registration (no:journalpost) is often used to register a document. This includes incoming, internal and outgoing documents. A document is associated with a registration within a case file

3. A Noark 4 database
This is the relational database that a Noark 4 systems stores data in. The standard specifies tables and integrity constraints. A Noark 4 database holds all the data products instantiated during the records management phase. The database represents the records management phase of Noark 4 records.

4. A Noark 4 extraction

[1] http://www.lovdata.no/for/sf/ku/xu-19981211-1193.html#map001

The Noark 4 database needs to be converted from data stored in tables to data stored in XML files. The extraction also include all the documents converted from production to archive formats. An extraction should be made only on casefiles that are finished. Any open case files should be move to the next period. The extraction represents the long term preservation phase of Noark 4 records.

5. Users during the records management phase

During the records management phase we identify four users (roles) that can be a judge of fitness of use. The first role is the case handler. The case handler is responsible for the case handling and will create registrations. The second role is the leader. The leader can assign cases to case handlers, approves outgoing documents. Both of these roles deal with case handling. The third role is the records manager. The records manager has the overall responsibility of the data in the Noark system. This includes the structure of the Noark database, periodisation, extraction generation. The records manager has the overall responsibility for the case handling. The fourth role is the municipality as a singular entity. The municipality manager can be identified as the person in the role and looks at the Noark system in terms of the legal requirements but also as flexible tool.

6. Users during the long term preservation phase

During the records management phase we can also identify four users (roles) that can be a judge of fitness of use. The first role is the person that processes extractions. This role will have the responsibility to make sure the archive has received all information that is required and will often have to check what is contained in the extraction. This ranges from the all case files, supporting data, to the set of documents. Everything should be checked and validated. Once this is done the archive can say it is in receipt of a valid deposit. The second role is that of a record locator. there will be many types of record locators. It could be someone who works in the archive or a member of the public looking for a particular record or someone locating records as part of a judicial case or a researcher looking at individual records as part of research. Either way they are looking for a single record in a myriad of records. The third role is a record processor. This is primarily a research role that is not concerned with individual records but at groups of records, processing them to understand more about what the municipality has done. Trend statistics is an example of the processing this user role would undertake. This role would often be software and not a person processing and will require data of high quality in order to be effective. We can also describe this role as being algorithmic as trend statistics might be just one of many different kinds of analysis that should be undertaken. The fourth role is the archive institution as a singular entity. The archive manager can be identified as the person in the role and looks at the Noark extractions in terms of the legal requirements of long term preservation on behalf of the municipality. In many ways the archive exists to preserve this information. This role will be concerned with cost and will probably aim to reduce the costs of maintaining extractions. Throughout the data quality literature cost is an issue and it is equally applicable here.

Noark 4 is a relational database based system and is suitable for automatic objective data quality analysis of data in the tables. In theory we would expect a high level of quality as the vendors have to achieve and maintain a certificate of compliance with standard.

As discussed in Deliverable 1, Noark data goes through three distinct phases. The records management phase, where the data instantiates through the daily activities of the municipality. The magnitude of this data slowly increases day by day. The second phase is quite short and results in a copy of the data in the Noark system being extracted to an XML based format where each Noark 4 table in the database corresponds to an XML file. The municipality deposits this XML extraction with a municipality archive that acts as custodian of the material.

Figure 3.1: *The three traditional phases data in a Noark system experiences*

The two phases of importance for data that will find itself within an archive are the records management and long term preservation phase. The extraction phase binds these two phases together and in many ways represents a signing off on the data giving access to the data to the archive. The traditional way of dealing with this is that the municipality deposits an extraction with an archival institution and formal ownership is not transferred until many years later. Data should be deposited at an archive after five years to ensure that a proper extraction has been created. Another twenty years go by before the archive institution formally takes over ownership of the data. The reason an extraction is created after 5 years is to ensure that the municipality is in fact able to create an extraction of the data from the system. If the archive was to wait twenty years, before getting any data it is likely that the there will be problems getting the data out due to technological obsolescence. During the twenty year period there are two copies of the data and further processing of the data can occur leaving both copies out of sync with each other. An example of this is the retention scheduling aspect of records management where records should be deleted after a certain time period. The deletion must be carried out on both copies, otherwise the data is not synchronised. The archival institution is not in possession of the original records management system, just an XML version of the data. While the municipality will often upgrade to a newer version or a competing product and the original data becomes part of an historical database that is often read only. However from what we gather this is more of a theoretical problem than a practical problem. It is unclear how many municipalities actively use retention schedules as most of the data has to be retained for documentation purposes.

Noark supports and encourages the use of periodisation, the ability to limit the accumulation of records within time periods. It is often argued that the period of data is gathered should coincide with municipality elections. When a new records management time period start it is often a good

idea to depositing a copy of the municipalities data at their local archive. This does not always happen so a municipality can see its database grow considerably in size before a deposit occurs. If problems occur when creating an extraction, the sheer volume of data may mean it is impossible to handle.

During the records management phase the data serves a purpose and that is to document the activities of the municipality. The long term preservation of the records is not necessarily in focus during this phase and the Noark system simply has to meet the needs of the day-to-day running of the municipality. It is worth noting that a municipality is not required to have an agreement with an archive and can instead preserve their own data, however this means that the municipality should ensure they have sufficient archive knowledge at both the theoretical and practical level. More and more municipalities enter inter municipality cooperatives to establish a shared municipality archive. It is within this context that we have our focus. It is hard to guarantee correctness when a municipality upgrades their Noark system converting the data within the earlier Noark systems to a historical Noark database. There are no commonly available tools that do this for Noark 4. If this upgrade/copy cycle is undertaken a number of times it it likely we will see problems with format idiosyncrasies and other problems due to evolving standards resulting in problems. The time scale we refer to here can be ten to thirty years.
A lot of the potential problems can be reduced by extracting to XML and putting the data in the Noark 4 system in a neutral format. The ability to measure the quality of an extraction is a desirable. Today there are limited options to achieve this and there has traditionally not been that much focus on quality in terms of Noark data.

Introduction to data quality:

The project has defined Data Quality in Report one as "t***o which degree data in a system is in compliance with the real scenario the data represents and is usable***". In Norwegian this is "***I hvilken grad data i et system er i overenstemelse med det virklige scenarioet dataen representerer og er brukbar***".

This definition has a similar meaning as the definitions used in the data quality research field. Traditional data quality is often limited to the records management phase of data and our understanding is that little work has been carried out on data quality from the long term perspective of data. One of the reasons for this can be attributed to the cost factor associated with bad data for commercial entities and that the archive field area is traditionally does have a reputation to embrace new technology.

The term "fitness of use" is common when it comes to data quality (Wang and Strong, 1996; Knight and Burn 2005; Fehrenbacher and Helfert 2012). The term is a little abstract as both *use* and *fitness* can be very subjective. But the term indirectly assumes the existence of a user that ultimately determines whether or not the data is fit for use. As discussed in report one archival

data has a long term perspective and it is difficult to identify the user and their requirements in the future. We can expect the use of automatic processing of the data in the future so a user may be software rather than a physical person. It can however be argued that a user is indirectly implied in terms of the type of data that has been captured.

Wang & Strongs definition on data quality is that it is *"data that are fit for use by data consumers" (*Wang and Strong, 1996). Wang & Strong also look at data as a product and argue that an information system (in our case Noark 4) is a data production system. In terms of the archival context and extraction of Noark 4 data the notion of data as a product is very fitting. Data is extracted from a from a Noark 4 production system and made available to an archive institution. This extraction consisting of XML files and original documents as PDF/A is clearly a data product.

A data quality analysis of a Noark 4 extraction should measure the quality of the data within the extraction. Data quality is often measured in terms of data quality dimensions where a data quality dimension is a collection of data quality properties that represent a single aspect data quality aspect. Completeness is an example of this and can measure whether or not a complete record has been captured.

Historically the data quality research field have identified three approaches to data quality measurements that include *intuitive*, *theoretical* and *empirical* approaches. The intuitive approach relies on the experience or intuitive understanding of which features are important for the data quality analysis. How accurate and reliable the data are is a good examples of this. A theoretical approach measures deficiencies intrinsic to a data product and derives data quality dimensions based on theoretical principles. Bad data quality here could for example be a result of deficiencies in the data production process. With an empirical approach, however, the focus is on the user of the data and capturing data quality properties that a user will likely deem as important.

When it comes to data quality an analogy of a production line of automobiles is often used. The product the end-user cares about is the car as a singularity, but from the manufacturing point of view, the car exists as an aggregation of multiple smaller parts. The manufacturer has to make sure that all the smaller products are of high quality so the end product is deemed to have high quality. The user of the car is not really concerned with the aggregation of all the smaller products.

If we bring this analogy back to an information system, we see that a data product is often created as a result of data entry and represents some real world entity. The classic data product example is *a customer* where it is imperative to have the correct name, updated address of a customer as well as history about interactions. Each customer entity is a product but the entire customer database can also be viewed as a product. A business will want its customer database to have a high a quality as possible, in much the same way a person buying a new car.

The analogy can also be applied to Noark 4 as an information system. It is sensible to ask what a data product is in terms of Noark 4. There are two obvious units that could answer such a question. The first is the document, whether it is an incoming, internal or outgoing document and the second is the case file. The incoming document will often result in the establishment of a case file and therefore is the starting point of a data product. The quality of the incoming document is outside the scope of the municipality so there is little the municipality can do about the quality of this document, whereas the quality of internal notes and the outgoing document is under the direct control of the municipality and can be controlled for quality.

From the archive point of view, the data product is both the entire extraction as well as the individual case file and documents. Some archive institutions have no desire to attempt to fix data quality arguing the responsibility is on the municipality to ensure the extraction is acceptable. Interestingly, one of the problems with extractions is the definition of what is acceptable. At one level an extraction can be deemed to be acceptable if it validates against associated DTD or XSD files. It is difficult to argue beyond this point but the arkn4 and URD tools described in Deliverable 1 go beyond validation tests and test for relational integrity and some other data quality aspects. It is unclear how much data quality we should test for before determining something is acceptable. It appears that different archive institutions vary with regards to what is acceptable and treat this very subjectively. As an outside observer, the logical approach is for the archive institution and the municipality to have joint responsibility for the creation of an extraction. However this could be a controversial point within the community.

When considering an extraction, in many ways the product is the entire extraction and it is essential that the entire extraction is deemed to have high quality. But when viewing an extraction as a data product it also consists of many sub-products. If we go back to the car production analogy, the municipality appears to have a focus on the individual components that make up a car while the archive is both concerned with individual components as well as the finished product.

There are no central definitions, reports or processes in Norway that can help understand data quality for this type of material. At meetings, the municipality archives tell that data quality is something they are concerned with but their focus is on preparedness to receive extraction from various electronic records systems. Data quality is an issue they will have to look at later but the need to understand data quality for archive material is rapidly increasing. A number of municipality archives have not yet received any Noark 4 extractions but in the coming years this will change and they will have to maintain and provide access to such material.

Data Quality Dimensions

Prof. Richard Wang of MIT is seen as one of the instigators of data quality as a research area. His early work is still relevant and today fifteen of the most common data quality dimensions are recognised. These are shown in Figure 3.2 and are grouped in four categories: Intrinsic,

Contextual, Representational and Accessibility.

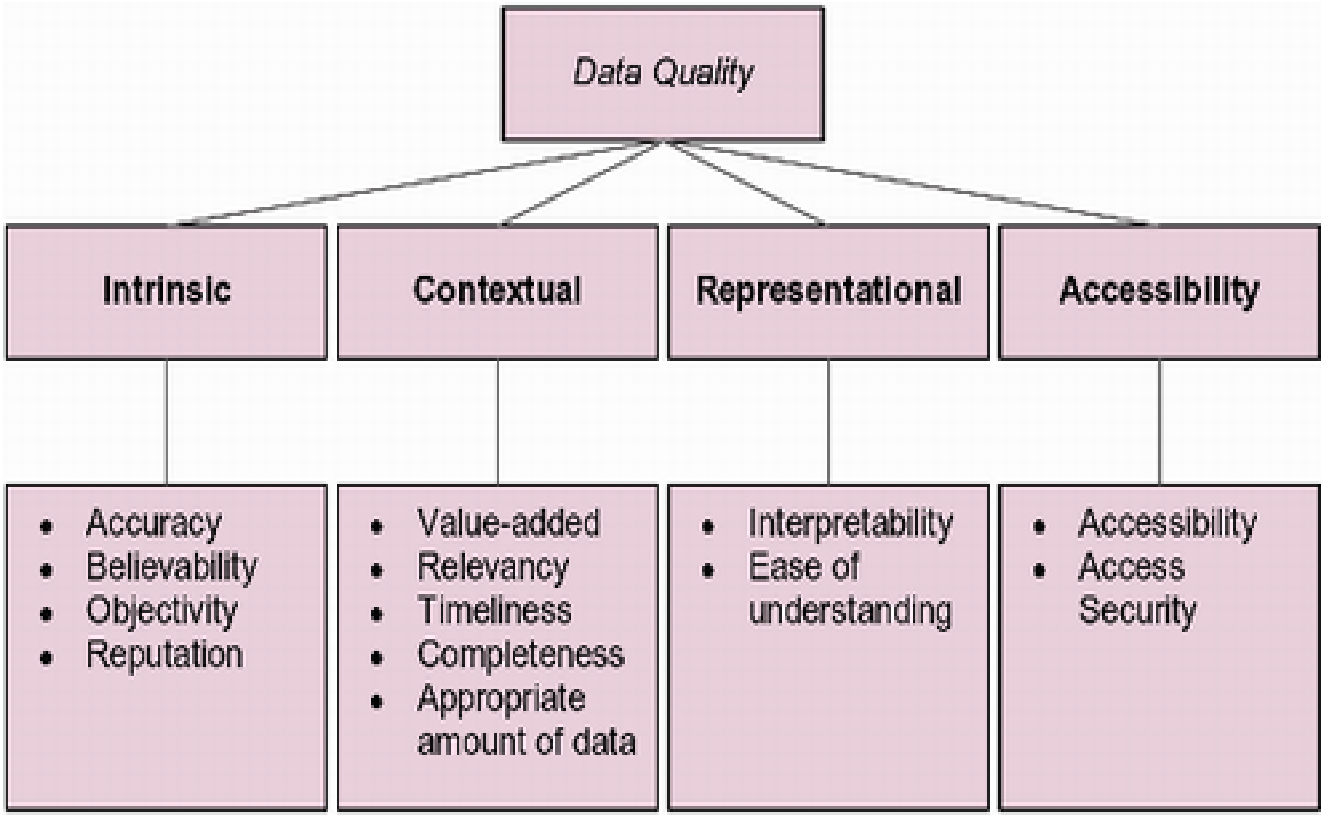

Figure 3.2: *Conceptual framework of data quality by Wang and Strong*

Intrinsic data quality

The dimensions that belong to intrinsic data quality are:

- Accuracy
- Believability
- Objectivity
- Reputation

Accuracy

The accuracy dimension reflects how well the data captured represents the real world object. It is something that can be measured objectively if information about the real world object exists independently. It is often denoted as the distance between ***v*** and **v'** where **v** is the data product and **v'** is a known correct representation of that data product.

Accuracy during the records management phase

A lot of the data sources within a Noark system can be checked for accuracy. Drop down lists that contain information are such a source. For example the basis with which a case is to be kept from being public always requires a basis in law. There are a set of common values used "§ 13 freedom of information act". If the system allows user to enter their own values then variations of the original values can quickly occur. The classification system being used and class values will often be a drop down list. These values can easily be checked for variations to check distance from **v** and **v'**.

In Noark systems a common real world object referred to in a document will be people and companies. In Norway the National Population Register[2] can say something about people. All citizens and residents in Norway have a social security number and are required by law to inform the National Population Register if they move to another location. There also exists a comprehensive company register called Brønnøysundregisteret[3]. During the records management phase there is scope to ensure that real world entities are correct in the Noark systems, but this does not always happen. In terms of **v** and **v'**, **v'** is the data product representing the company details in the Brønnøysundregisteret while **v** is the data product representing the company details in the case handling system. A distance function like Levenshtein could easily be used to measure this.

The use of social security numbers has historically not been common when dealing with municipalities unless it is an important aspect of a case. A kindergarten application for example needs the social security number of the child and parents while a social security number is not that important for building application. Often the social security number will not be registered in metadata but be part of the incoming document or application.

The current trend is that more and more interaction with governance takes place in an online settings and eventually we will be forced to use our social security number to login via the altinn service when contacting the municipality for standardised services. A positive side effect of this should be increased accuracy in the ace handling systems as the chance for errors in data entry are reduced.

But this is also something that can be measured subjectively. For example an address register of addresses that the municipality is frequently in contact with. A change in contact person or address that a case handler is aware of might not be updated in the system. In this case it can be difficult to ensure there is an external register that we can ensure contact information is update municipality.

We have used people or companies as an example of the accuracy dimension, but the person or company will have identified themselves in an incoming document. So if the company has misspelled their name or address it has a relatively small consequence. Often the postal service can handle this. The Norwegian postal service also has a list of valid addresses in Norway. A

[2] http://www.skatteetaten.no/no/person/folkeregister/

[3] http://www.brreg.no/

misspelled email address could be a data quality problem if the reply address is not the same as the incoming email address.

All roles during this phase should strive to achieve the highest accuracy possible, however factors like time and pressure to finish a case and move on to the next case may mean that accuracy could suffer. Of the four roles it is the case handler that undertakes most of the data entry tasks and in many regards we expect the case handler is responsible for most of the accuracy. Over time substandard work may be too difficult to fix due to volume. The leader role often reviews case before it is finished and can approve or send a case back for further processing. In larger organisations some case handlers have the authority to sign off certain types on their own case files on behalf of the leader.
The records manager has an overall responsibility of accuracy in both the handling of case files but also the archive structure and to ensure all data sources used are accurate.
We can also distinguish here between accuracy in the Noark system and accuracy in the documents. It would be relatively straightforward to measure accuracy for a lot of the data sources within a Noark system, the archive structure. **v** and **v'** are in many cases obtainable but in terms of the documents we do not have a **v'**.

Accuracy during the long term preservation phase

In many regards the accuracy of data products will fall as the number of data products increase over time. This will be as a result of a level of expected fuzziness as people with similar names enter the archive. For example according the national statistics office 167 men[4] have the name "Marius Olsen" in Norway and in total there are 50655 people that share the Olsen surname. An archive at national level will have trouble differentiating between all these Olsen people. The only distinguishing factor will be their address, but unless the archive has access to the history of peoples addresses to disambiguate then fuzziness is to be expected. The archives inherit a data quality issue due to the increasing size of data and for every new data product received, we can expect to see a drop in accuracy at the collection (all extractions) level, while the accuracy at the case level will not be affected. In terms of **v** and **v'** as the collection of records grow the usefulness of **v'** diminishes unless something is done to ensure we can uniquely identify **v**.

Ideally extractions are not edited once accepted at the archive. Updates should not happen to the original object but should be applied to a derivative of the original object. If a record later shows itself to be incorrect, for example a registration is put in the wrong case file, then the original case files should be left as they are and new derivative objects should be created that shows the correct state. Standard like PREMIS use this thinking.
In terms of **v** anv **v'**. The fault (registrations are in wrong case files) that is detected results in a new **v'** and when **v'** is compared to **v** a distance is detected. Updating the v so that it is accurate with **v'** fixes the incorrect state. This approach might be useful in terms of ensuring integrity when updates to extractions are required. To achieve this the archive needs to be in a position to identify **v'**. A similar approach applies to handling deletions that result from the enforcement of retention schedules. The data will be deleted and the system maintaining the data has to

[4] http://www.ssb.no/navn/

register that the deletion has taken place.

When an extraction is deposited and processed a check for accuracy should be undertaken. Some analysis of what the extraction is should be undertaken, especially if the extraction contains data with retention scheduling.

For a record locator accuracy will be important. Finding aids like classification codes and variations of spelling of names can be important. For example the Haugland name has a variation Hauglann which appears over time from a common ancestor. It is not uncommon for the Hauglann name to be misspelt as Haugland. A better understanding of a person via altinn service mitigates this risk, increasing accuracy.

The record processor role will be severely limited in terms of functionality unless the extractions have high degree of accuracy built in. A person locating a record can use intuition and common sense. This does not yet exist in analysis of electronic records.

For the archival institutions extractions from various Noark systems from various municipalities over varying time frames will challenge to accuracy, especially across a collection. Norwegian has two official Norwegian language forms (bokmål and nynorsk) and the use of dialects will be cumbersome for automatic accuracy measurements on Norwegian language extractions. In addition there are a few minority languages with varying official support. Even though there are a lot of information retrieval techniques that can be used to aid the situation, the archive must strive to ensure as high a degree of accuracy as possible. Accuracy is a potential cost issue for an archive.

Accuracy and Noark 4

Accuracy is a dimension that is relevant to Noark 4 for objective measurement. We have to first identify data products and sources that have the potential to determine a **v** and **v'**. People and companies are not relevant to this measurement. We do not have access to v' for people. For companies we do have access to information about the company via Brønnøysundregisteret but unless the company Identifier is we can not guarantee v'. Internal to a Noark 4 we have a number of list-like data sources like classification system and laws. These can be used to identify **v** an **v'**.

Believability

The believability dimension reflects how well the data captured is true and credible. Most of us have at some stage received an email from a relation to a prince fleeing a war torn land and the promise of a share of millions of dollars if we allow that person to borrow our bank account. Often a small upfront fee is required. For many people the mail is not credible and most people know the the story is a lie. Yet people still fall for these scams.

Believability is also something that applies to records in a Noark system. In terms of a record that is used as part of a judicial case a judge may have to consider the believability of a

document. We can not assume that just because a document exists in the Noark database that it belongs there. It has to be associated with a registration, which in turn is associated with a case file that belongs to the archive structure. If these are not in place believability can be questioned. Similarly if a document has disappeared from the Noark database but there is an empty registration then we may believe the document existed. Believability is related to provenance and is at the core of record management and archival principles. If a document from a Noark system is to be objectively believable then the provenance of the document from its creation to today has to be documented. Of all the data quality dimensions believability is probably the most important one to capture and arguably might be the single most important data quality dimension for Noark 4.

Believability during the records management phase

When it comes to believability of Noark data products we can distinguish between believability at the content level and believability at the object level. Content can be exemplified as the information relating to a case, the written words in an application for child assistance, the name of a parent or child. This is information that is typically held within the document or application form. At the object level we mean believability of the record as it exists in the Noark system, that the document the case handler sees on the screen is the same letter that the sender sent, that the case was created on the date stated in the record.

An example of a record where credibility could come into play is for example a person applying for child benefits for sixteen children. The average number of children per family in Norway is between two and three. Whether or not this information is true would have to be verified by checking against the real world entities, the children. In this case the credibility could be questionable as it falls so far outside what is considered the norm for Norway. It does not however mean the record is not true.

An example of a record where truth, but not credibility comes into play is for example a person applying for child benefits for 4 children. The person may not have four children, that is there is no truth behind the data in the record. Distinguishing between truth and credibility here, it is credible that a person could have four children.

The above examples are not that far from reality. In 2010 the media reported[5] on a case before the courts where people had tried to trick the welfare service claiming benefits for children that did not exist. As a result twenty-seven fictive people were removed from the national people register. In this example believability could be seen as subjective. In Norway the people register exists to ensure that this kind of fraud does not take place. Hospitals report when a child is born and the child is registered in the national database.

There is a distinction between believability for the various roles. The case handler may have to make decisions with regards to believability based on the contents of the case and experience in case handling may play an important role here. The leader may also have to make a believability judgement on the contents but will not have the same amount of time as the case handler. The records manager will probably differ with regards to believability with a

[5] http://www.dn.no/forsiden/politikkSamfunn/article1962324.ece

primary concern about believability at the object level and accept believability of case handling. When creating an extraction the records manager will aim to ensure the extraction is believable and pass it to the archive is such a state. Similarly the municipality will be concerned about the reputation of its records and the way it undertakes its case handling. In some regards believability is related to other dimensions. A Noark 4 extraction may not be believable if it does not completely and accurately represent the Noark 4 database.

Believability during the long term preservation phase

The processor will not have time to consider believability of individual cases and documents. Neither is it something the processor or the archive wishes to do. The processor will need to process the Noark 4 extraction in such a manner that whatever believability the extraction had during the records management phase is maintained during long term preservation. The processor may question believability if there are obvious shortcomings. This could be that the extraction is missing important files like the casefile table or some if not all the documents are missing. Elements of this dimension exists in archive theory as fixity. Fixity allows us to objectively define believability at the object level (the entire Noark 4 extraction) but does not give credibility at the content level.

Record locators that are researchers will often have the ability to put the records they are viewing in context and it is often something seen as subjective based on training and experience. Believability will be one of many factors they consider when they locate records. If the record processor is software then believability is not necessarily an issue. The person that starts the analysis may or may not be concerned with believability. If the goal is trends statistics then minor errors will not have any significant impact when processing large amounts of records. The analysis might even be a believability analysis comparing known trusted sources to sources where trust is unknown.

Believability and Noark 4

Believability when it comes to an extraction is not a binary issue, it is not as simple as you believe or you do not. Rather it will be a decision that is reached depending on a number of factors. Are all files that should be in the extraction present, do the numbers add up - is the number of cases per year plausible? Believability can be assessed through the evaluation of evidence. A Noark 4 extraction consists of a number of XML files that contain a copy of the tables from a Noark 4 database. A file called NOARKIH.XML is created that states the number of tuples from each table. If this file is missing, then it could be an indication of some other problem. For example the extraction process could have crashed resulting in the file not being created or the developer created NOARKIH.XML file after the extraction of individual tables takes place and counts records from the XML file instead of the source database. These kind of systematic faults could severely impede the believability of an extraction. The programs that undertake such extractions are not subjected to scrutiny and the only requirement in the Noark 4 standard about this is that NOARKIH.XML must contain information about the extracted tables. However we assume the software vendors do this correctly.

For the most part it will be difficult to measure believability at the content level, especially in terms of the contents of documents. Information at the object level, relating (metadata) to the documents can to a degree be subjected to measurement. It is possible to check for completeness, that all metadata elements are accounted for and where possible measure accuracy. Dates should be valid and sensible, i.e the creation date is plausible given other factors (start and end dates given nearby records). When considering the entire Noark 4 extraction a checksum can provide believability that the extraction from creation to delivery is unaltered. There is little need when considering an extraction level to require proof that the the municipality did indeed create the extraction, a quick look at the contents extraction will reveal this.

Some parts of believability can be measured objectively. The use of checksum, a believability analysis etc can be employed for this task. From the point of view of a Norak 4 extraction, it might be easier to measure believability objectively than subjectively. The amount of data products in an extraction means that the archive really can not say anything about believability, they simply can not process it manually.

Objectivity

The objectivity dimension reflects how accurate, concise and unbiased the data is.

Objectivity during the records management phase

The case handler will be responsible for entering data and has the responsibility to ensure objectivity and experience and available time at hand will often have an impact on how objective the case handler is. There is already a focus on this during this phase and examples[6] of rules governing these kinds of fields can be found via the arkivplan website. Such rules are useful and when applied can mitigate the effects of bad objectivity due to the inherent natural subjectivity of various case handlers.

An example where objectivity is required is the case file title. It should be accurate and complete and should include the type of case and what the case is about. For example in the case of a company complaining to the municipality that someone working for the municipality damaged their property the title of such a case should include:

- the type of case - “complaint”
- what the case is about - “damage of property”
- who the case concerns “The Apple Import Company”

A similar argument applies to the title (*Innholdsbeskrivelse*) for a document as part of a registration. Here the title should be used so as to easily identify the document and but also has secondary purpose to separate the documents in the case file from each other. Both of these examples show how accuracy and conciseness should be used to ensure objectivity. An

[6] http://kommune.arkivplan.no/content/view/full/151648

example of how unbiasedness could be an issue is where a case handler is asked to tone down the title of a case or document in a case that could be viewed as volatile. Objectivity should also be prevalent in the documents. A case handler should not have any prejudices when handling a case.

From the leaders point of view, the leader will have the responsibility to ensure that the case has been handled objectively and will have to approve the outgoing document. The case handler will often create data but the leader should ensure objectivity. The archive leader will have the final responsibility but their focus will probably be at a more abstract level to ensure the Noark database has objectivity. The database should be accurate, concise and unbiased. In many ways the Noark 4 standard ensures this. The underlying relational model with ACID (Atomicity, Consistency, Isolation, Durability) ensure objectivity and a Noark system should not have bias

Objectivity during the long term preservation phase

Both the processor and archive institution wish to preserve truth and as such will not have any desire to update or fix objectivity. The case handling that took place is documented and will exist as evidence of the events that took place in the archive. The lack of objectivity at the content level is to be preserved and perhaps analysed, not fixed. It is difficult for the processor to show bias in handling the Noark 4 extraction, bias could be shown by the processor in terms of description but the extraction should not be touched other than to fix mistakes.

The records locator or researcher may have objectivity as a factor when collecting individual records. Objectivity is a subject that is often of interest by reasearchers working with archive data. One of researchers in the archive groups at HiOA wants to study prejudice in case files identify descriptions of people who collaborated with the Germans and who had children with German fathers. These children were called “Tyskerbarn” and reflects a shameful era in recent Norwegian history. This work is a good example of the records locator and manual records processor role in one. Collections of records will processed, but individual records will be very important and analysed. As more and more records are electronic we open up for wonderful research opportunities in the future. The technology for automatic objectivity/subjectivity analysis through the use of language analysis in documents incorporating linguistics and semantics exists.

Objectivity and Noark 4

Our understanding through the existence of guidelines and from feedback is that there has been a long standing focus on objectivity when it comes to these fields. The Noark 4 standard sets the maximum size of the title field to 255 characters. This allows for approximately 40 words including spaces. The URD software actually does a test for number of characters in these fields with a stopword list removing words like “About” etc.

When it comes to the documents contained within a registration an analysis could be undertaken with regards to the objectivity of the contents of the document. This would involve linguistic and semantic processing and is beyond the scope of this project as it would be a data

quality analysis of the case handling process.

From the archives point of view accuracy, conciseness and unbiasedness take on a whole new meaning when viewing an extraction. The archive has to maintain the extraction in the form it was created. The objectivity of the extraction can be seen as the extraction being synchronised with the the contents of the data during the records management phase. When it comes to a level of unbiasedness in the extraction, the extraction must not be carried out in such a way that some data products are missing or in an incorrect state. The archive really has no interest in the objectivity of individual case files. The case files are what they are, they reflect the actual chain of events that took place.

Reputation

Reputation as a data quality dimension refers to the degree with which a source or contents of an information source has a high degree of reputation often reflecting the trustworthiness and importance of a given information source.

Reputation during the records management phase

It is during this phase that many data sources are used, and the different sources will have varying reputation. A case handler will naturally use a number of data sources while undertaking their tasks. These can be anything from lists of values to choose from to actual sources of explanations of law and guidelines for handling certain types of cases. From the case handler perspective an example of reputation could also be related to timeliness. Consider an internal resource that is related to an external resource. If the external resource changes then the internal resource also need to be updated. When a case handler wishes to keep a case or parts of a case out the public domain a paragraph of some Norwegian law must be applied. This is often in form of a list of laws that can be applied. So when the law is updated the internal representation of these laws must also be changed. If this does not happen in a timely manner, it may have an effect on reputation.
Another example of this dimension is also prevalent in some institutions in the third level sector where student addresses are often stored in multiple systems but case handlers intuitively know that one system will have the correct term address student. One source has a better reputation than other sources of information. In this regard we can see link between reputation and believability.
When considering a specialised system (*no:fagsystem*) that is integrated with a Noark 4 system. The reputation of this system as a source can be positive or negative based on how the user experiences data being passed between the systems. In this regard we can see link between reputation and believability. The archive leader can start to question the trustworthiness of the data in the Noark system if problems are prevalent with the integration of other systems.
The reputation of the case handling process is also interesting, as some cases may actually affect the reputation of the municipality. A controversial building application could be a typical example if a particular developer is seen to have been given preferential treatment. In this example one case file of potentially many hundreds of thousands can have an extreme negative affect. Interestingly this shows that reputation from the municipalities point of view can be based

on peaks rather than not based on average and that over time bad reputation dribbles off. Also considering the effect of openness in terms of Noark data. The more openness you have, there is a potential for better reputation but more openness could also lead to a negative reputation it the public are not happy with what they see.

Reputation during the long term preservation phase

In this phase data sources are not used, but the Noark 4 extraction in its entirety becomes a data source and will acquire a level of reputation Specialised systems can have a bad reputation with regards to their extraction because the systems were never designed with extractions in mind. So it is not unlikely that these extractions will have a bad reputation. When it comes to Noark 4 systems, there are seventeen different Noark 4 systems with certificates of conformance. Some of these systems have evolved into newer systems and typically have regular updates. It is difficult to gauge which system and version is in use across all the 419 municipalities and the county and state levels. Each of these systems should produce Noark 4 extractions with varying degrees of quality. They should all produce 100% correct extractions but we know that is not the case. These systems will naturally develop reputations. We know now that some municipalities have had problems producing complete extractions from the system at the centre of the run in Report 3 but there is no overview of what the situation is like for all systems. An interesting question we should ask is whether or not the ability to produce a complete extraction is a system or data problem or combinations of both.

The Noark 4 extraction processor in the archive will have certain expectations about the level of difficulty in processing an extraction based on the reputation of the system and over time will get to know the systems and the their quirks. Similarly the researcher looking for individual records in the future will probably find common traits across extractions that come from similar systems, and reputation might mean that a given extraction is not seen as easy to work with or believable. When it comes to automatic processing quirks in the various extractions may mean that the automatic processing software avoids reputations from certain systems. From the archive institutions point of view they may gain a reputation for having good or bad competencies in dealing with such extractions.

Indirectly the use of data sources with a low reputation during the records management phase should reduce the reputation of the Noark 4 extraction. However, this extremely difficult to capture as it would require that an understanding of the reputation of the sources used are also passed on to the archive. As such the use of data sources with poor reputation will for most cases have little or no consequence in the archive. This could be dealt with by

Reputation and Noark 4

The fact that Noark 4 exists and is in use gives the data the municipality creates a certain reputation. If the standard had not existed municipalities would have used various systems and we would probably see varying degrees of quality on the data. We can argue this based on the diversity with the specialised systems and how in the General Auditors report from 2010 states that no specialised system have met the minimum case handling requirements and that of 666 specialised systems in use in the municipality of Oslo, only a few them have some form

extraction ability. The use of Noark gives reputation to the data and to the institutions that have used systems based on the standard. Another role that is prevalent in the reputation dimension is that of the software provider. The inability to create correct Noark 4 extractions may have an impact on the company's reputation.

Contextual data quality

With contextual data quality, the users focus within a context is the central focus.

The dimensions that belong to contextual data quality include:

- Value-added
- Relevancy
- Timeliness
- Completeness
- Appropriate amount of data.

Value-added

The Value-added dimension focuses on what added value the data gives the user and questions whether the data is beneficial and provides advantage to the user.
If we look at the reason a municipality uses a Noark 4 it is because it has to document its activities, there is a legal requirement that Noark is used and value-added should be evaluated in this regard. From the archival perspective what added value do we see on a set of extractions? A lot of municipality archives are in the process of preparing to receive deposits of Noark 4 and Noark 5 extractions. A lot of these institutions need updated skills and competencies. Initially these extractions will challenge the archives to update their skills and competencies leading to the archival institution gaining a stronger reputation with regards to handling electronic material.

Value-added during the records management phase

Does a Noark system add value beyond the legal requirements? If we look at this from a user perspective, a case handler, leader or archive leader we could get varying degrees what added value each experiences. An earlier study we undertook at HiOA as part of a masters identified negative feelings of the the Noark system. The study observed that the case handlers wished far more flexibility with regards to viewing and sorting and processing case files. A common perception is that a Noark system meets a requirement and has little focus on the user.
From the municipalities point of view what added value does a Noark 4 system bring beyond the legal requirements. This is a question a municipality manager could answer and we would expect the answer to be that a Noark system helps with an efficient and effective running of the municipality.
A Noark system also provides the records manager with a timeliness in terms of functionality provided from the records management system by allowing for periodisation, updates and changes to classification system and other data sources. These make the job of records manager easier.

Value-added during the long term preservation phase

It can be argued that Noark 4 extractions gives value in terms of providing a complete picture of the data and documents for the archive. Value-added for the extraction processor is a

standardised approach to dealing with an extraction and various Noark 4 systems should really produce structurally similar extractions. For the archive manager, reduced costs for long term preservation and the re-use of tools. For the record locator value-added can be in the form of being able to retrieve and find records in a timely manner.

Value-added and Noark 4

Perhaps the value-added from the noark standard is one to society in terms of increased documentation control and reduced records management operational costs. The points in time when extractions should be deposited is not a requirement of Noark. Noark is a system, requirement. The fact that the points in time time when deposits happens varies is a consequence of other factors, and can actually be a result of other data quality dimensions. For example do the archie institutions have a good reputation in terms of handling an extraction. A bad reputation might mean that the municipality will prefer the maintain the data themselves.

Relevancy

The relevancy dimension captures whether or not information can be seen as applicable and helpful.

Relevancy during the records management phase

The municipalities have a legal requirement to document their transactions so all data in a Noark 4 system is relevant. Junk mail incorrectly inserted and test data perhaps are an exception to this. One of the reasons to use a Noark system is to capture relevant information and to keep irrelevant information out of the system. With regards to helpful, the Noark system documents the municipalities activities and provide access to records. In that regard it has to be helpful. Perhaps by reversing the question, we could attempt to see if the Noark system is *unhelpful*. In some regards we hear from users that the systems can interrupt their workflow, they spend a lot of their time outside of the system doing case handling work. An argument here can be made that Noark 4 is mainly about documentation but some users expect more from the system.
For the case handler the system is clearly relevant as the documentation of the case handling is done here. The same can be said for the leader as through the system outgoing letter are approved. The records manager will find it useful as it gives them the tool to exercise control and verify that things are undertaken correctly. The same can be said for the municipality.

Relevancy during the long term preservation phase

For the archive a Noark extraction will also be seen as relevant. The long term preservation and access to archive material will be central to the archives mission statement. All archive data that falls within the archives mission statement will be applicable. Most valid and correct data will be helpful to archive, increasing its collection size. When it comes to bad data, there could be a increased cost issue associated with preservation even though the data has to be preserved. If the archive is ready ready to process and maintain extractions then relevancy will be an issue. Tools and knowledge are central elements here. An underfunded archive may have problems

with extractions and they will not be seen as helpful. The extraction is relevant both for the extraction processor and the archive manager. For the record locater extractions will be relevant if they contain the data being searched for.

Relevancy and Noark 4

Noark 4 is extremely relevant for municipalities as creates a system to ensure that relevant information can be captured. Processes employed in the municipality must ensure that relevant information is captured, while a Noark 4 systems sets in place a mechanism to capture the data.

Timeliness

The timeliness dimension captures whether or not the data is sufficiently up to date. A common saying about this dimension is “Data delayed is data denied” . If data is required within a given time frame and the data is unavailable within the timeframe then the data is unavailable. Depending on the context the consequences of denied data could be anything from mild to severe. Examples of this dimension are up to date information about customer details, address, email etc.

Timeliness during the records management phase

In terms of a Noark system, timeliness with regards to contact information should not really be an issue. When a case file is created it often includes the contact information of the people referred to in the case. Even if the address of a person changes later, the address was probably correct at the time a letter was sent. Timeliness can be an issue with contact information of external entities. But as we saw the Brønnøysundregisteret holds up to date information on companies that can be verified. Timeliness can be a minor issue with regards to zipcodes. Infrequently zip codes are updated and this could cause some data quality issues unless the source adequately gets this data updated in a timely fashion.
The quality of timeliness will vary from setting to setting. When it comes to search, we expect a result in under a second and anything beyond that is often not deemed slow. We can view the same effect in a noark system. If the system is slow in presenting data the user can have a negative perception of the system. We know of one Noark 4 system that warns the user if they are about issue a search the “*” operator that issuing the query might reduce the system performance experience for other users.
An example where timeliness in terms of case handling has a financial cost is if the municipality exceeds the time requirements for handling a building application. The municipality will need to ensure that processing is handled efficiently so that they don’t miss out on this important funding source. § 7-4 of the building regulations[7] state that a municipality must reimburse the developer 25% of the case handling fee for building projects for each week or part thereof the time limit of handling a building application is exceeded.
For the case handler, timeliness can be related to information sources

Data quality dimensions related to time are often categorised as belonging to process dimension. While it is clear that this time dimension associated with Noark data during the records management stage has a financial cost, it has no financial implication in terms of long

[7] http://www.lovdata.no/for/sf/kr/tr-20100326-0488-009.html

term preservation.

Timeliness during the long term preservation phase

From the archive perspective timeliness can be whether or not the municipality is timely with depositing material. A major problem for municipality archives can be that with migration to full electronic records management, the archive will find itself in a position that it suddenly has to receive many terabytes or archive data. Timeliness for an archive with regards to extractions is to receive data on a continual basis avoid peaks and the inevitable queue of data to be processed.
When the extraction is created the data lose the natural ability within the system to be updated. A certain dynamism to the data falls away and the data becomes more static. This is not the case if a deletion/retention schedule kicks in.

For the record locater role access and understanding can come into play under timeliness. In terms of access, if the data is stored on a tertiary storage then delayed access to data could be seen as data being denied. In terms of problems with the extraction generation, that it is not possible to create an extraction that can be validated then it is important that the extraction that can not be validated be preserved as it is a correct description of the extraction at that time.

Timeliness and Noark 4

Timeliness is a property associated with data and the entity that data represents and as such is not a Noark 4 issue.

Completeness

The completeness dimension is concerned with the data products having a wholeness for the task at hand, that is there are no data missing and breadth and depth of information is appropriate. Fields that are defined as optional and compulsory often make up the analysis of this dimension.

Fields that are compulsory must contain a value. The problem with compulsory fields is that people sometimes enter random or rubbish data to meet the compulsory requirement. If the person entering that data believes the compulsory field is not relevant to the task at hand, it is easy to enter false data. Not everybody has an email address, even though the majority of people in Norway have an email address. It would therefore be wrong to require an email address for each person in a case handling system.

Completeness during the records management phase

Completeness is definetly central to the case handlers work. Some Noark systems provide drop down lists so a case handler can easily insert required information. A Noark system could prevent a casefile from being created if the required fields are missing. Sometimes however the required information is not available. i.e. an email address or phone number is missing or illegible and some effort has already been put into getting the case file registered. In these situations the system becomes an annoyance for the user and the user often forgets about the importance of gathering the information. Sometimes a user enters dummy data, i.e many

spaces or question marks just to finish the process they are undertaking. The records manager will have a responsibility to ensure the casefile is complete, that an obvious missing document is found and associated with the case file etc. The records manager will also wish to believe that at a collection level, all records are complete and there is no missing data. This is unlikely to be achieved however. The municipality will also wish for the same, but as they are a level higher than the records manager it might be harder for the municipality to grasp why everything is not complete.

Completeness during the long term preservation phase

The extraction processor will be very concerned with completeness and for an extraction will undertake a number of tests to ensure that extraction is complete. The archive manager will be concerned with completeness to ensure that the archive has all the collections they should have and that no information is missing from any of the collections. The municipalities vary with regards to the number of extractions they have deposited with a municipality archive and if most of municipalities have deposited a copy then the archive manager might wish to ensure the collection is complete and ensure the remaining municipalities deposit a copy. The records locater will definitely be concerned with completeness. Once a record or set of records is found, they should be complete and no information is missing. The algorithmic role will probably not be as concerned and just look for whatever it can find.

Completeness and Noark 4

Noark 4 does define what completeness from a Noark perspective by defining mandatory and optional fields. Also the standard often defines default values.

Appropriate amount of data.

This dimension is a reflection on whether or not the volume of data are appropriate to the task at hand.

Appropriate amount of data during the records management phase

We have to distinguish between a data product as what we see on the screen (i.e. a casefile) and the Noark system as a data product. For the case handler and the leader role the system should present an appropriate amount of data for the task at hand, there should be no information overload or cramming of information. For the records manager an appropriate amount of data might mean too much data. It is important however to distinguish between the amount of data personnel resources required.
From the municipality perspective, the data in a Noark 4 system is probably appropriate as volume is not really an issue as the database fills up gradually day by day. Every day new case files are created and documents uploaded or produced. If the ability to fix bad data quality is an issue the municipal is concerned with then a Noark 4 system as a data product quickly becomes an inappropriate amount of data.

Appropriate amount of data during the long term preservation phase

Information Technology has advanced such that a municipality archive should not have any trouble storing the amount of electronic records the municipalities produce. Preserving them such that they are understandable in the future is a much more difficult task. There is a point at which an extraction becomes difficult to process due to volume. Time is often an issue here as a lot of time is spent copying and processing large extractions. An extraction can easily be anything from a few GB to a few hundred GB. When the extractions become very large there is an inappropriate amount of data to process at a single moment in time. However the actual amount of is probably appropriate if we look at the requirement to preserve records from a Noark system. For the extraction processor a lot of time can be spent setting up environments for the extraction. When validating a large extraction it could actually take over a week. This means the extraction processor has to work in parallel with other tasks. From the archive point of view, can there ever be an inappropriate amount of data to preserve? There is no explosion in the amount of data a municipality generates it is in fact controlled to a degree by the rules governing what should go into the Noark system. The use of specialised systems are a major problem for archives and these can have a problem in terms appropriate amount of data. The record locater role can have problems with appropriate amount of data as collections that are large and/or heterogenous in nature might be too difficult to work with. The algorithmic role will probably be developed to work with large collections of data.

Appropriate amount of data and Noark 4

Noark in many ways limits the amount of data to an acceptable level. There is often argument thrown around that the amount of data we generate increase exponentially and that we will not be in a position to find individual data products in the future because we will drown in information. the problem with that argument is that it excludes the development of new retrieval tools but when it comes to municipalities, there is no data time bomb. Through the use of a Noark system the data that has to be reserved is already captured and dealt with. Yes technological changes are a challenge, the use of social media etc. But the standard indirectly defines an appropriate amount of data.

Representational

The dimensions that belong to representational data quality are:

- Ease of understanding
- Interpretability
- Consistent Representation
- Concise Representation

Ease of understanding

First we clarify an issue between understanding and interpretability. These these two words are often confused. See the the section on interpretability below for more information. This section is concerned with how easy it is to comprehend the data. You have to be able to understand something before you can use it. Take for example a text written in another language you might be able to understand that there are words and sentences but you might not be able to interpret them. Similarly data written in in various Norwegian dialects might be difficult to interpret.

Ease of understanding during the records management phase

We have to distinguish between the system, metadata and content in when looking at ease of understanding. The system can have an influence in how the user understands data. A badly designed system with bad process rules or a bad user interface can affect on understanding of data. When it comes to metadata it should be clear what the various metadata elements are and mean. Take a fictional metadata object “Casefile date”. Is this the date the casefile was created, last accessed or closed? There should not be ambiguity. Similarly the same can be applied to the documents. How simple is it to understand the answer a case handler has written.

Ease of understanding during the long term preservation phase

Understanding an extraction can be a major problem for the record processor. To a certain degree the amount of documentation associated with the extraction and the level of experience working with extractions will have an influence on this. For the archive manager understanding can be both a cost issue as well as a preservation issue. It is often argued that to achieve long term preservation the following three constraints have to be in place. First you have to be able to store the information, then you have to be able to read it and finally you have to be able to understand it. To be able to preserve with understanding in mind the archive will seek out file formats suitable for long term preservation, but they have to address the storing and reading issues first.
The record locator will be very concerned with understanding records and depending on the level they have access to information, understanding can be a challenge. If XML files are the only source they can access then they first have to understand XML.

Ease of understanding and Noark 4

The standard in many defines a lowest level understanding for how records management should be undertaken and provides a common platform for all interested parties, the people purchasing a system, designing a system and using such a system

Interpretability

The interpretability dimension captures whether or not the data product is understood. The main difference between *interpret* and *understand* is that you can understand something without being able to draw a conclusion. When you interpret something you can often draw a conclusion. Often the words understand and interpret are used to mean the same thing but there is a subtle difference. One can also argue that to be able to interpret something you first have to be in a position to understand it. In the previous section we presented how easy it is to understand a data product while in this section we discuss if the data object can be interpreted. The ability to interpret a data product is based on the data product having meaning. We can not understand a data product unless we are in a position to interpret it.

To achieve a high level interpretability definitions within the data product are clear and an appropriate language is used.

Interpretability during the records management phase

From the case handler perspective interpretability can be a system issue or a semantic one. A case handler can have data in front of them and know that it is a Noark case but has difficulty interpreting what the case is about.

Interpretability during the long term preservation phase

From the extraction processor point of view interpretability can be seen at the system level. To be in a position to to create a Noark 4 extraction you need to understand the Noark 4 standard and the underlying database. To understand how the database maps to the Noark 4 standard, you have to first interpret both the database structure, tables and attributes and the Noark standard. A successful interpretation is required to be able to create an extraction. The same argument can be applied when dealing with an incoming extraction. You can understand that a set of files is an extraction but you need more knowledge to interpret and use the extraction.

Interpretability and Noark 4

The standard as documentation ensure interpretability. It provides answers to structure and gives meaning to metadata elements. Records Management as defined through the Noark standard defines data products very clearly. Both the Registration/Document relation and case file data products clearly define the data products.

Consistent Representation

Symmetry and consistency are two traits we like when processing information. When it comes to consistency we look for data to be presented in the same format. An example of consistent representation that could cause problems could for example an American entity communicating

with Norwegian public office. The american data format is MMDDYYYY while the Norwegian format is DDMMYYYY. If the date is an important aspect to registered as part of a case then an inexperienced case handler may easily convert the month to day and day to month. An example is the date 03042012. Is this the 4th March or 3rd April? A similar example can be seen with the use of metric versus imperial units.

In a Noark setting there is an established consistency of how casefiles are represented. This is clear when looking at how case files and registrations are represented, the two representations are interrelated. Within a single Noark 4 system this should pretty much be consistent but within the archive we may see varying use of this.

The official format used for dates in Noark 4 is YYYYMMDD. The screen format can differ from the database format. When creating an extraction this date format often requires a processing from the standard database date formats with the potential for error in processing.

Consistent Representation during the records management phase

For the case handler and the leader a consistent representation can be how metadata objects are described within the user interface. This could also relate to terminology used within the case handling area and the use of synonyms. If this is achieved, there is a potential for less confusion. For the records manager and the municipality roles within documents . Across systems, for example between the case handling system and specialised systems it is possible the systems and user interface use synonyms describing the same or different entities. This can lead to confusion and bad data quality.

Consistent Representation during the long term preservation phase

For the extraction processor, the contents of metadata and documents will not be of interest so this role will not really concern itself with the consistent representation. This role may see a lot more problems with consistency between system extractions. Municipal archives will naturally process extractions from various Noark 4 systems and there really should be consistency between them in terms of naming conventions. Dates will follow a fixed format, but there might not be consistency in terms of identifiers like casefile or regististrations. The use of synonyms for similar or dissimilar entities can be a problem for long term preservation. If looking at this as a semantic issue the archive manager will probably not be too concerned, but at a system level issue and costs associated with long term preservation the archive may find it necessary to undertake a disambiguation and documentation process to ensure this is not a long term preservation problem.
The record locator might experience consistency in the same way as extraction processor. This role may also find the lack of consistency a problem and may have trouble locating records because of it. The algorithmic processor will also find consistency difficult to deal with unless a disambiguation process is undertaken.

Consistent Representation and Noark 4

Consistent representation is not part of the Noark 4 standard and really something that the data producers are in charge of.

Concise Representation

In terms of a concise representation we define it as the degree to which the data product is compactly represented.Concise implies that a description is brief, but includes the relevant information. Consider the following two examples for a case file tittle.
"*This case is about the damage that took place on the companies wall where the wall was damaged allegedly by the one of the municipalities blue vans delivering food to the elderly*"
When the title should really be more concise "*Complaint about damage to "The 'Apple Import Company' wall by municipality van*". This dimension is not something that is defined within the Noark 4 standard, but something we see for example from arkivplan.no with regards to prodeures for describing titles.

Concise Representation during the records management phase

The aspect of concise is related to data entry and as such is a semantic issue belonging to the people dealing with data entry, the case handlers and the leaders. This dimension is heavily set on these two roles. The archive manager will have the final responsibility to ensure that a concise representation has been used. Concise in this scenario can be also seen as being related to objectivity and we will probably wish that metadata elements that are to be described are concise. Can we apply the same argument to the documents the case handlers work on, replying to citizens or the documents presented to the politicians? In some cases there will be a wish for conciseness and in others there will be a desire for more information. An example of this is where the Norwegian social welfare and work office (NAV) started a process in 2012 to use a simpler language when communicating with their users. Central to this process is to avoid long sentences and excess text. Concise is a nice word to describe this.

Concise Representation during the long term preservation phase

The extraction processor will probably not have any particular opinion on the conciseness of the records but will process them accordingly. If we were to extend the definition of concise beyond data and metadata to a system level, concise can be interpreted to mean that the extraction should be small but including the relevant information. That is an issue the extraction processor is concerned with. A similar argument can be used for the archive manager role. The records locator will have an opinion on the conciseness of the data products. Depending on the information need and time available, conciseness will be an issue. If the record locator is trawling through thousands of records looking for information then conciseness will probably be appreciated. It is hard to tell how the algorithmic role will look at this. We have come a long way with natural language processing so this might not be an issue for this role.

Concise Representation and Noark 4

Conciseness is not part of the Noark 4 standard and really something that the data producers are in charge of.

Accessibility

The dimensions that belong to accessibility data quality are:

- Accessibility
- Access Security

Accessibility

In terms of accessibility it is assumed the information is available and easily retrievable.

Accessibility during the records management phase

From the point of view of the case handler and the leader accessibility is often a system issue. A user is granted access to the system and the data products within. The Noark 4 system is there to ensure accessibility. It is nearly easier to visualise this if we reverse the situation and argue what the situation would be like if there was no Noark 4 system and the case handlers were working on documents located on their machine and using email to communicate with the leader then we can quickly see information might not be available and easily retrievable. Versions of documents would be a very good example of this and data loss through editing of various versions could easily occur. Noark puts in place a system to handle this. Noark brings traceability to the process and ensures that data products are accessible to both the archive leader and municipality.

Accessibility during the long term preservation phase

The migration strategy takes the data out of the system and puts it into a neutral long term preservation format. Without a retrieval system around this information accessibility can quickly become a problem in the archive. As it stands at the moment the data is only stored in xml files and there is no fulltext search on documents. This is certainly true for what we have seen in the municipality archives but it could be different for the archives at state level. Search in an extraction is then limited to opening up XML files and using the [CTRL]+[f] function in an editor. For the records processor accessibility also includes the need to ensure no information is encrypted or if it is to ensure that it is decrypted in the archive. No municipal archives have the infrastructure to deal with production format encryption and it would be a nightmare to maintain such a system. The records processor has the responsibility to ensure accessibility for the data products are maintained for the future.

Accessibility will be important for both the record locator and algorithmic role. Can we expect the record locator to be able to work with XML files to search for material. Although some users probably would be able to search through XML files, there will be a need for a retrieval solution. Without such a solution we can argue that accessibility is not possible. In the near future this is not a problem as only the archive will be able to search the material but in 60 years when the data is made available then it should be possible. We have to envisage that within the next 60 years the technology to handle these kind of heterogeneous data sources with various data quality could be available. Having the data in XML based formats is a really good situation for this role so with the data out of a system and in a neutral format really makes the accessibility possible for this role. From the point of view of the archive manager accessibility will be related to its reputation and existential issues. It is imperative for an archive institution to ensure its

collections are accessible. In many ways the use of archive formats and the migration strategy ensure accessibility.

Accessibility and Noark 4

Noark 4 is heavily tied to the relational model where the underlying database is the foundation for many applications. This is very much in keeping with the way the software industry worked at that time and the approach the Noark 4 standard was timely and relevant. The intentions behind the standard were not met by the all the software vendors and we have a situation today where for some systems it is difficult to create an extraction and ensure accessibility to data. But the reliance on the data model and the use of technical specification mean that accessibility to a degree is possible. Retrieval is at the heart of the Noark 4 standard via SQL. the choice of technolgy has been crucial in ensuring accessibility. In the long term perspective we need documents i neutral formats and archive systems that allow access to material. In some cases material that it is not expected to be used very often can be stored on tertiary storage with an access time from minutes to hours or days. As the data volume for the archive increases the archive may be forced to store some data in slower systems or on tape. A National archive will over time easily have to store many petabytes of information and from an economical point of view it may not make sense to make all material retrievable in seconds. Usage patterns can determine this.

Access Security

When it comes to access security the aim to ensure data products are appropriately restricted to ensure to maintain security. Any record manager would be appalled if documents in their Noark 4 system were to be leaked, but there seems to certain types of documents where the records manager would be more appalled. Child welfare cases where the social services have been involved are often deemed as the most sensitive. In some situations these documents are encrypted within the records management system. As an observer, this is a peculiarity as the if you encrypt documents within the records management system, you have a problem with access control. In many regards there is a fundamental flaw with regards to trust.

In terms of an extraction of sensitive data archive can have encryption but that is a barrier to access. Access Security has to be at a system level and the object on terms of the data products. This domain is also an extremely important domain for both records management and archives. We have to take into account Authentication and Authorisation. Authentication means you have been identified correctly while authorisation is whether or not you have access to a particular data product.

Wikileaks is a good example of data quality and access security. Some of the material reported in the media from wikileaks points to a problem about access security. Whether the issue is a trust issue or a system issue is debatable. A system should not really allow a user to copy large amounts of data. An example of this is that military installations sometimes have two computers for employees. One of the networks is totally locked down where USB ports and other IO devices are soldered shut so data can not be copied. This does not however preclude someone from taking pictures of what they see on the screen. Access security ultimately boils down to

trust.

Access Security during the records management phase

During the records management phase access security will be regulated through the use of user identity, passwords and general IT security. The standard sets the model for access security and additional IT security can be used to ensure physical access to the system. There is probably a greater risk to access security if the case handler works outside the constraints of the system or brings documents home. The new Bring Your Own Device (BYOD) strategy being employed where the enterprise allows users to bring untrusted devices into the enterprise network will also be an issue here. It will become more important to confine documents within the boundaries of the system to ensure secrets are kept secret. The municipality role has to balance access security and the modern way of working. The archive leader can and probably will be a bit more pedantic with regards to access control. The leader and case handler are really limited to the access security as defined within the system and probably will reflect too much over it. We have been told that sometimes there is a type of shadow case handling that can take place, where case handlers have documents and handle cases outside of the system. We have never seen this but have heard from practitioners that this situation can occur. If some people work outside the confines of the system then the access security takes on a whole new dimension is weakened.

Access Security during the long term preservation phase

The records processor will not have too much focus on access security as it is assumed it should fall naturally within the security restrictions the archive already has in place. Procedures and systems should be set up to ensure that no accidental or malicious access or copying of data should be able to occur. For example a processing area for incoming material could be on a machine that has no USB/DVD writing capabilities or Internet connectivity. To protect the extraction processor these kinds of restrictions should be in place so that the person can not be accused of leaking information. The archive manager will be very focused on access security as it will be central to its reputation as an archive. For this role the archive has to balance accessibility against access security. There will be a wish to publish as much as possible, to make as much information as possible accessible and at the same time only make what should be made available accessible. Authentication and authorisation will be important system aspects. Physical access will of course be a major issue for the archive manager.
The record locator may or may not be concerned with access security, when locating records this role will wish to ensure that all records have been located. While satisfying an information need, this role might find references to other records that have restricted access and will wish delve deeper looking and asking for access to records until the information need is satisfied. Access security can be seen as a hinder to achieving the successful location of records. The algorithmic role will find this more problematic. It will not necessarily see the potential of related records and notice that not all records have been located. This role will approach this is a more binary fashion. It simply has access or it does not.

Access Security and Noark 4

The standard defines users and roles and has access security as one of its core issues. This is only relevant during the records management phase as once the system is migrated to an XML format access security as defined within Noark 4 is no longer relevant. The user in the future will be the archive, not the case handler, leader in the municipality. The distinction between the long term preservation and the records management phase really become apparent when considering access security.A lot of effort is put into ensuring access security during the records management phase but once the extraction is made, all that security really boils down to being documented security. Within this context, the standard shows itself as being more a records management standard and not necessarily a long term preservation standard. That is probably as it is meant to be, it is a records management standard with an eye to long term preservation.

References:

1. Fehrenbacher, Dennis and Helfert, Markus (2012) "Contextual Factors Influencing Perceived Importance and Trade-offs of Information Quality," Communications of the Association for Information Systems: Vol. 30, Article 8.
2. Knight, S. and J. Burn (2005) “Developing a Framework for Assessing Information Quality on the World Wide Web”, Informing Science (8)1, pp. 159-172.
3. Wang, R.Y. and D.M. Strong (1996) “Beyond accuracy: What Data Quality Means to Data Consumers”, Journal of Management Information System (12)4, pp. 5-34.

4. Data Quality Run (Deliverable 3)

In Deliverable 1 and 2 we defined what data quality is and applied the general approach to data quality to Noark 4. In this report we apply this knowledge to a Noark 4 extraction scenario and undertake a data quality analysis on Noark 4. In some cases the presented results will be very specific to a given system, while in other cases the results will based on observations and discussions with the field.

Background Information

All municipalities in Norway that are part of a municipality archive agreement will at some stage deposit their electronic data with a municipality archive. Few municipalities have done this for various reasons. There does seem to be an accepted urgency with regards to this issue for some municipalities, while others seem not to be that concerned. One reason for this situation could be the lack of understanding of the implications of long term preservation of data in record management and specialised systems. There is various support for extraction tools among the vendors and the lack of extraction tools is inhibiting an efficient and effective solution to the problem. Pricing is also a major issue as some municipalities do not have a budget to carry out this kind of work. It has been argued during discussions that if this job was put under the IT budget then the IT departments would be a lot quicker and more interested in finding a solution than they are at the moment.
One of the researchers at HiOA was interested in learning more about Noark 4 and data quality but privacy concerns laid down in law have impeded that work. The group has been working on developing open source tools for records management and archives. The reason for this is the closed nature of the systems today and the General Auditors report shows that the municipalities are currently facing big problems, especially when it comes to handling electronic material. The closed nature of the systems inhibit education of a new type of records manager/archivist that also understand the technical aspects of systems they are guardians of. This project is a part of a number of records management and archives projects at HiOA that approaches RM/Archives from an integrated life-cycle perspective. The researcher has a good understanding of Noark 5 and his skill sets included programing and database administration.
In order to learn more about Noark 4 and what kind of issues municipalities and archives face with this electronic material the researcher agreed to develop a Noark 4 extraction tool and create a Noark 4 extraction for a set of municipalities using a particular Noark 4 system. This system has a built in extraction tool that can but the common understanding is that this tool does not produce an extraction that can be independently verified with for example *arkn4* or *URD*. We did not get a chance to verify this as none of the municipalities had a license to use the built-in extraction tool.
To the best of our knowledge this is the first project that is based on data quality working within municipal electronic records context and as such we wish to avoid giving an impression that a certain municipality has a records management system with bad data quality. Remember the litmus test for data quality is based on the *user* and as such we cannot simply quantify a data quality value. Such an analysis could give a very wrong impression of a municipality or a given system if we were at this early stage point at data

quality shortcomings. Therefore the municipalities in question and the Noark 4 system will not be identified. All descriptions of the system i.e. tables names etc have been changed. The person that developed the code and performed an extraction is defined as a data handler for the municipalities and has signed confidentiality agreements. We stress that **no research** has been carried out on the data. Any results presented here are presented from an analysis of the code that was developed and the accompanying public reports that describe what problems were fixed and from discussions with various relevant people during the project.
In many ways the extraction process that the developer undertook describes the situation where an archive has been given a database dump with no documentation and has to undertake preservation efforts on the database. Understanding Noark 4 and the types of data quality issues that can arrive in this scenario is very important. The related project on developing this extraction tool and how to fix bad data attached to a Noark 4 system provides useful insights for archives and record managers. Working with data pre-extraction also makes the results of this project more applicable to specialised systems.
An important issue to note is that the extraction was created from a copy of the database. A backup of the database is created and saved as a single file. This backup file is then read and the database can be installed on another computer. The extraction software did not make use of the actual database connected to the Noark 4 system. When we report on data problems with the database we report on problems with the version that is a backup. There could be some additional functionality or otherwise that we never got to see. We do however believe that the backup accurately reflects the data in the running system as the backup was a full backup. This is another reason why we feel it could be misleading by identifying the system. *It is also worth noting that the tool developed by Frode Kirkholt known as URD can, for some Noark systems, act as an extraction tool but it did not support the system at hand.*

The extraction tool that was developed here used a basic Noark 4 database structure with referential integrity switched on. The reason for this is that it was initially assumed that the database would not follow the Noark 4 standard and that there should be some data quality issues. The approach we took was to map the loose interpretation of Noark 4 to an actual Noark 4 standard so that we could create an extraction.
The data quality work is a combination of syntactic, semantic and pragmatic analysis of working with Noark 4 extractions. Our observations are used to generalise about the data quality issues that can occur with a Noark 4 database when creating an extraction. What is interesting about this approach is that we get to work with the data at the point it is to be extracted. Some systems will fix issues before they create the extraction and others will simply generate a set of XML files that can not be independently validated for correctness.

Data Quality and Compliance with Standard

Very early on it became clear that compliance with the standard was an issue. Table names did not follow the standard. In some cases it was possible to identify the correct Noark table as table name was contained within a more complicate dtable name. For example the source database could name a table called JPOST as ZXSEJO. This mapping is documented somewhere but not publicly available.

But is the lack of correct table name a non-compliance issue with the standard? That question nearly becomes philosophical in nature but is easily answered with a No. The standard itself does not require strict adherence to tablenames or otherwise. It begs the question of why bother with Part 2 of the Noark 4 standard. The point of part 2 with table structure is to serve as a guide to how the modules described in Part 1 of Noark 4 could be implemented. If the goal is to be able to create a valid extraction that can be validated with a tool like arkn4 then you need to have the data in structure similar to how it is described in Noark 4 Part 2.

When the National Archive develop a standard like Noark, there is really a potential benefit for achieving higher data quality if comparing against a situation where no standard is used. Given that the software vendors can implement a database structure as they see fit then some of the potential DQ benefits from using Noark 4 quickly fall away. Strict compliance to the standard would have resulted in a potential positive gain with regards to data quality. It is also known that some vendors have committed themselves to a stricter implementation of the standard than others and we would not be surprised if we in the future see better data quality in extractions from systems with stricter adherence to the standard.

Aside from naming conventions, relational dependencies and the use or lack of use of primary keys are issues that potentially could result in bad data quality. We did find these issues in the database and in many ways it was surprising as one would not expect it in professional grade software.

As the vendor can freely interpret the standard we took a closer look at the columns in the tables and found the column lengths also deviated from the standard. This was done when mapping the database back to a Noark 4 structure and all column definitions were documented. We found a number of deviations from the standard both in length and type. The type deviations were not really an issued but the length deviations have a potential to cause interoperability problems. For example one column was defined in the Noark 4 standard as X(10), while the Noark 4 database implemented the column as VARCHAR (15). We checked and note that none of the values in the column have a length over 10 character so it was not really a problem. This type of issue can be detected with the URD tool.

One of the issues we ran into is that our knowledge of Noark 4 was not as deep as what we first assumed it was. A number of people working in the field agreed that in some cases the various roles from records management to long term preservation have a shallow knowledge of the Noark standards. The knowledge is often enough to get by on daily tasks but quickly becomes a problem when technical questions are raised. So our own knowledge of Noark 4 itself becomes a potential hinder to good data quality. This has a potential to become a problem for the municipal archives in the future. As time goes by knowledge about Noark 3 and Noark 4 will naturally leave the archive as people come and go so it is important that this knowledge is not allowed to fall away. Nobody in the project group was an expert in Noark 4 although there was various levels of knowledge about records management and extractions. This combination of knowledge gave us an interesting starting point for the

project from a long term preservation perspective. How could an archive institution develop a Noark 4 extraction from a database dump? What kind of data quality issues would be faced? The project would ultimately result in deeper knowledge of Noark 4 for all the institutions as well a better understanding of the challenges these database dump file pose and how to approach the data quality issue with them.

Our understanding of Noark 4 was limited to the technical section known as Part 2. As such our comments on the standard should be seen in that light. The standard is in some cases underdocumented and in one or two instances has misleading information with regards to referential integrity between tables. The under documentation issue leads to holes in knowledge with regards to what a field represents. This becomes an issue when a vendor does not use the same names for attributes or tables as the standard does. In one or two instances we believe we found the standard says table A links to table B, when it fact it refers to table C. Given the 95 tables and nearly 885 columns, this kind of issue can naturally happen.

The arkn4 tool does include an example extraction but this is not adequate to fully understand how to create an extraction. The example extraction of arkn4 is probably something that is left over from the time the tool was developed and the example extraction is extremely useful when trying to understand values in a Noark 4 database. It is clear that there should be more example extractions (with fake data) that the R&D community has access to as we do not have access to data like an archive institution has.

However the documentation is an adequate source for someone with a records management and database development background. But it did become clear during the project that the level of documentation in the standard is very important issue.

Data Quality Analysis

In this section we present an overview of the data quality analysis that we carried out. Time was one of the major issues we faced and this overview really just scrapes the surface in terms of what data quality and Noark 4 is. This analysis is based on tables and data in tables. Dealing with files is an issue on its own and became so big that we assigned an own section to rather than it being an under section.

Object Status

In our case the municipality purchased a new system and were left with the data in the old system. The new system did get a copy of a lot of the records from the old system as a historical database. The need to close everything correctly in the old database falls away in this situation. In fact it can be cost prohibitive to manually close all unclosed case files etc. Object Status for a case file should at extraction stage be “A” or “U”. There will however be a lot of unclosed case files in an extraction. This is an issue that we have been told by many people working in the field is something that is very common in municipalities in Norway and a clear DQ issue. In one case we heard of a municipality with over 2000 unclosed case files that were being manually closed through the GUI. The records manager was simply opening casefiles and closing them and this is a job that could just as easily be done using a SQL command. arkn4 and URD have the ability to detect for these kinds of values and report on them.

NULL Values

NULL in a database has a special meaning, that a field has not been assigned a value. It is central to the relational model that NULL means that a field is unassigned. If we find NULL values in a Noark 4 database, how should we interpret them? We really should interpret as not being assigned a value. An example of this issue iw where the field SA.U1 in a casefile is used to denote whether or not part or the whole title in a given casefile is public or not. A value of 1 means the casefile title is not public, a value of 0 denotes the casefile title is public. The database however uses a NULL value to denote Noark 4 value of 0 and a value 1 to denote a Noark 4 value of 1. In many ways this kind of analysis will be perceived as pedantic but without access to documentation to ensure this situation is correct the extraction software could inadvertently undertake a formal evaluation of the public situation for a casefile .
Another example is for the table SAKTYPE. Here there is a required field called KLAGEADG and again should contain either a 1 or a 0. However all instances of this field were NULL.There are not that many records in the table and as such it is a lot more difficult to determine whether or not values have been set or not. Should we interpret NULL as 0 or is data actually missing.

Missing Information

A lack of documentation is a problem when working with a Noark 4 database, especially when the database does not follow the naming conventions of the standard. This leaves many issues open to interpretation and missing information is often frustrating and time consuming to resolve. Two examples of missing information are presented here. The first is tables that are missing, the second is values that are missing. A number of tables were

missing. The Noark 4 system did not employ all the modules available from the Noark 4 standard so it is not clear what tables should or should not be present. An example of two of the tables were missing are LAGRFORM, LAGRENHET and NUMSER. NUMSER we were unable to figure out and will engage in further study, while LAGRFORM is a list of the file formats in use in the extraction and should all be long term preservation formats. We could not locate this information anywhere in the database and instead had to manually construct the file. Similarly LAGRENHET is a file that defines unit for storage. It was not clear how it should be used but we were able to manually create such a file. This kind of issue is really frustrating when working with extractions as so much time is spent investigating.
It can be argued however this really is showing our lack of understanding the documentation. We believe there are very few people with this much technical knowledge in about Noark 4.

In terms of missing data we found examples of where data that is required to be part of the production system and as such the extraction was missing. The table TGMEDLEM details which user is a member of which group and identifies the user that added this person to the group. The column INNMAV holds the identifier of the person that undertakes this action. From a provenance perspective at system level this can be seen as being important. However in nearly all tuples in the table TGMEDLEM, the INNMAV is NULL. We know that only one of two users undertook this action in the system. This missing information does appear to be a system problem and we hypothesise that at some stage an update to the system resulted in this formation being lost. To get around the problem a FIXER user is created so that the values no longer are NULL.

Column Dependencies

Column dependencies between columns mean that values in one column determine the value in another. For example for case files there dependency between SA.TGKODE and SA.UOFF. If SA.TGKODE contains a value other than the value “XX” then SA.UOFF must contain a description of why the details from the case file are not public. There are a number of functional dependencies throughout the standard.

Redundant and Additional Information

An extraction can contain information that is redundant. For example not all classification codes might have been used but these codes are in the table (ORDNVERD). Does this additional information create an uncertainty that there could be some missing data? Not all codes will have been used so it is natural to expect that there will be some redundant codes. But if a common code that it is expected that all municipalities use is missing then there could potentially be a problem. The example of classification codes is one that is often cited during discussions and there is an argument that all classification codes in the system should be part of the extraction and not just the ones that were actually used.The reason for this is that the codes can give an important insight into all functions that municipality was capable of carrying out during the time period of the extraction.To further problematise this issue, consider a municiaplity where a specialised system is used instead of a Noark 4 records management system for a particular function. In this case the Noark 4 system will have no references to the function via a classification code while all the information is in the specialised system.This kind of issue makes it harder to process and work with these sources in the future.

Another relevant situation is where a municipality has temporarily used a Noark 4 module to test it out. Shortly afterwards the municipality decides that the use of this module is to be discontinued and only a few records are found in the tables associated with the module. This leads to questionability in terms of the records found. Are the records real records or just test data? Should this data be included in the extraction? From a cost perspective how do we balance the cost of writing, testing and verifying the code for this module, especially given the fact that one of the tables are not in compliance with the standard and some manual processing will be required to wash the data so it is valid. Can we categorise this information as redundant information? If the data is more administrative in nature and does not fall under registration obligations laws then this information may be deemed redundant from a Noark 4 perspective. It could however fall under the definition of having archival value.
Redundant Information would be considered unnecessary and it is within the context of Noark 4 we need to define what is necessary or not. There is definitely a certain amount of information that can be categorised as additional in a Noark database. This information will be unique per Noark system is and vary with regards to the system vendor. This kind of information will be very important from a system/daily use perspective but plays no formal role in a long term perspective and is ignored completely in the extraction. The type of information here can be system tables, search logs, access logs, system update log, usernames and passwords etc. Here we need to make a formal decision on what is relevant from a long term perspective. The structure of a Noark 4 extraction is a formalised decision on what has to be preserved but search and access logs could just as well be important sources of knowledge that is worthwhile preserving. The system update log can contain references to various scripts that changed the database structure and/or updated values. This issue is dealt with in more detail in Chapter 5 when comparing the data in the system to the data in the extraction from a DQ perspective.

Referential Integrity

Referential Integrity is at the heart of the relational model and really is an understood concept when modelling databases. An example of where referential integrity is used in Noark 4 is between a CaseFile and a Registration. There should exist no registrations in the Registration table unless they are linked to a case in the CaseFile table. Using referential integrity with cascading in this case will ensure that the database will never allow a registration to be created without an existing casefile or if a casefile with registrations is deleted, all corresponding registrations are deleted to avoid orphans from occurring.

The system we worked with had not switched referential integrity on. We assume the application is aware of the database structure and updates tables appropriately when users enter, modify or delete data. In many ways the situation begs the question, “Why would you not use referential integrity?”. The answer could be as simple as to be able to handle updates to the system from user requirement changes and so on or it could be a result of upgrading the system/database from an earlier standard i.e Noark 3.

Some relationships between tables are mandatory, while others are optional. For example the relationship between a registration and a casefile is mandatory in Noark 4, while the relationship between a series and a classification system is optional.

The use of referential integrity between tables with a mandatory relationship can mitigate potential data quality problems. In our own work with the extraction we did not get as far as testing whether or not referential integrity at the casefile/registration level was an issue, but the issue did manifest itself between other tables. For example there is a relationship between the Series table (ARKIVDEL) and the Classification system table (ORDNPRI) table via ARKIVDEL.PRIMNOK and ORDNPRI.ORDNPRI. This is an optional relationship and as such a series does not need to have a classification system associated with it. However we found that in a number of Noark 4 databases when looking at the records associated with that series, that these records did in fact use a classification system. The lack of referential integrity posed mild problems with regards to creating an extraction.

Primary Keys

Primary keys are used to uniquely identify a record in a table. For example a casefile will have a unique identifier to differentiate the various casefiles in the system. The system we worked with did not use a primary key in any of the tables we examined. Instead it designated the columns that contained what one naturally denote the primary key as being “NOT NULL”. This allows for the potential occurrence of duplicate records and in fact was one of the issues we came across a number of times when creating the extraction.
An example of this problem was seen in the table ADRTYPE that had a duplicated row that was common across all databases and as such believe this is probably a configuration/ installation problem. To remove the duplicated row is a minor issue, it is a simple delete statement, but it increases the time taken and complexity and as such increases the cost of creating an extraction.
In another instance and one that is far more difficult to fix is where the table DOKTYPE had duplicate primary key values on ND.DOKTYPE with the value “I”. The first was "I” for incoming letter and the second was I for what we believe to be an incoming job application. We had not tested it but others have commented that these values are used in different series (arkivdel) objects and as such can probably be distinguished in an extraction. The probable solution here is that the second “I” used for job applications will be changed to a different value an all corresponding fields will be updated
The third example of this issue is where the primary key used in the table has to be identified as a combination of different fields in the table. Take for example the Noark 4 table SAKART, The primary key is called ST.TYPE(10). The database allowed for a replication of the ST.TYPE value six times. However we were able to use an additional field in the table and were then able to get unique values. However these values are referenced in other tables and we then have to update the values in the other table as well.
The non-use of primary keys is noted as a serious inhibitor from a creation of extractions point of view and a negative data quality issue.

Case Sensitivity

Case sensitivity can also be a problem. This is especially a database issue. We observed an example of this is with status values. A particular status value can be declared twice, once in upper-case and once in lower-case. The reason for the two status values is unsure but it could be as a result of a mistake by the records manager or due to a software update.

An example of this is illustrated with data taken from the delivery status table (forsendelsestatus). The data in this table is extracted into an XML file called FSTATUS. XML. The following two XML snippets show this issue.

```
 <FSTATUS>
    <FS.STATUS>L</FS.STATUS>
    <FS.BETEGN>Levert</FS.BETEGN>
  </FSTATUS>
```

XML Snippet 1

```
 <FSTATUS>
    <FS.STATUS>l</FS.STATUS>
    <FS.BETEGN>Levert</FS.BETEGN>
  </FSTATUS>
```

XML Snippet 2

The difference between the two code snippets is that snippet 1 uses a capital L, while snippet 2 uses a lowercase l. With a case insensitive database the addition of the second "l" results in an exception (The record already exists) while a new record is added in a case-sensitive database. The Noark 4 standard shows that this value should use a capital L. If the problem was limited to database additions it could easily be dealt with but if there is a dependency on this data then the duplicates could cause problems. Especially if the table that uses this data also contains duplicates. This is probably not a problem during the records management phase of the records lifecycle but when it comes to creating a valid Noark 4 extraction problems can ensue. Semantically this is not really a problem, but syntactically this is a big issue. In this case the data was edited and all references to the little "l" were replaced with a large "L".
This problem could manifest itself semantically if the "l" and "L" had a different meaning, that "L" stands for "Levert" while "l" stands for something else. Washing the data in this case is less obvious.
This is really a problem when it comes to testing the extraction using arkn4 or URD, not when the data is in production. But given that vendors update their systems regularly the problem can manifest itself in various forms and become difficult to fix.
The simplest solution to handling this problem is to make the database case-insensitive but that might cause some unforeseen problems.

User Interface Issues

We also came across an issue where it was known that the user interface of a Noark 4 system had problems with classification codes known K-Koder. There is dependency between these values within these codes and the use of one value can require the use of another. The system was in use for period of time with this problem and it is expected that this error will work its way into the archive. This is problematic from a search perspective where searching in the future might be done on the K-Koder and if these are wring a user may not find a correct grouping of records.

Summary

A lot of the issues here are classical data quality issues at the syntactic and semantic level. Some of them are actually just database design issues. The issues here can be measured objectively, that is no matter how many times we measure or analyse for these problems we will get the same answer. While working with these issues we became aware that. We did not have time to work with semantic issues at the content level but note the functionality in URD that analyses the length of title fields for case files and registrations.

Data Quality and Extraction Creation

How the database is processed to develop the extraction can be an extremely important factor. A Noark 4 extraction is really just a list of the various Noark 4 tables in XML format. However if the vendor does not implement the standard strictly and parts of the Noark 4 standard is realised through the user interface then it is not possible to simply copy the data from the tables to an XML file. As with the system we worked with, one is left with an understanding that Noark 4 is realised as a combination of a loose interpretation of the standard and the application.

Two approaches can be used when extracting data from a Noark 4 database. The first is a simple table-copy approach, the second is a more complex traverse-structure approach. With regards to the first approach, if the database is a strict interpretation of the standard, it should be straightforward to simply copy data from the table to an XML file. However care must be taken to ensure that there is no additional functionality hidden in the values of the table. This could for example be that a casefile was deleted as it was never meant to be created. This casefile could still exist in the table but be simply be marked as deleted with a field that is not part of the standard. A simple copying of the table data could result in this casefile being restored in the extraction. This issue becomes even more complex if the vendor has a loose interpretation of the standard. The second approach is based on trying to understand the structure of the archive by traversing its formal structure. With this approach the code starts at the highest level (fonds) and works its way down through the archive structure to the lowest level (document). This approach can force the person creating the extraction code to identify and fix problems. Both approaches ultimately result in the table based extraction and have pros and cons. The approach we applied to our extraction code is a hybrid approach. In most cases a simple table extraction is enough to create many of the XML files, while in other cases a more complex approach based on an understanding the archive structure is required.

From discussions it is clear that the various Noark 4 systems have issues with regards to creating valid extractions. There are so many reasons why this occurs and some are reflected in our data quality analysis earlier. A reflection around this issue is that perhaps a greater focus should be given to this issue rather the bells and whistles of the front end. We are left with an impression that for some vendors maintaining the code for extractions is an afterthought. Some record managers informed us that they have been told that a vendor will make an excuse that the person who wrote the extraction code no longer works for the company and as such patience is expected. From a coding point of view it seems that, for some of the systems, as the systems evolve over time, the database structure and functionality changes and is put through QA, but the extraction tool does not necessarily follow the evolution or is integrated into QA process.

Data Quality and Files

One of main approaches anyone working with extractions will undertake is to first locate the files. Files can be located inside the database or outside the database in a specific folder. In fact the Noark 4 system we undertook the extraction code for used a combination of both approaches.

Storing files in a database has some potential loss problems, especially if converting the database from one DBMS to another. For the files that were located within the database, they were stored in a binary format in chunks of approx. 32 000 bytes. So if a file is 77 020 bytes large it will be spread across 3 tuples in the table with an identifier for each row on how to put the file back together again. At the time the system was developed this might have made a lot of sense as a lot of the files were probably not that large. As time went by the average file size has increased and it is not uncommon to have to a large number of PDF files that are over 100MB large in the database. From a performance point of view it can be ineffective to store such a large file in so many small chunks but from a preservation point of view it is irrelevant how the files are stored. The main thing is that the files can be retrieved. What is more worrying from a long term perspective is that the datatype used for this information is not a universal database data type i.e it is not available for example in MySQL. As such the data has to be converted from one datatype to another if the data is being imported into MySQL for example. Again when looking at this issue one must consider the design from a production Noark 4 system, not from an archive point of view. Some archive institutions actively use database conversion software but this type of software is not always mature enough to handle all of these kinds of cases and there is a potential for data loss when transferring a database from a DBMS that uses non universal data types to another DBMS.

There is a common discussion as whether it is better to store the files in the database or outside of the database on the filesystem. When the data is stored in the database it is potentially harder to access the files to change information, compared to storing the files outside of the database on the filesystem. But that argument is not necessarily true as it really boils down to trust and ultimately the Noark 4 extraction has no information about who accessed the Noark 4 database, when they accessed it and what they changed. Adding the files to the database increases the size of the database dump files, that is a potential problem from a processing point of view as the time to work with dump files increases. The databases that have been at the centre of the reports we base this report on vary in size from 18GB to 138GB.

Traditionally the municipality archives do not have large IT budgets and have often used standard desktop machines for this kind of work. The size of the databases do have an effect on data quality in the processing dimensions but can also also have a consequence on cost as a lot of time can be spent working with the databases. The import command in DBMS we used spends a considerable amount of time importing a backup file with all data and files into the local system. To handle all databases dump files, production copy of files and archive formats of files we had trouble fitting everything within a 1TB disk.

We also note that it is common when working with Noark 4 extractions that when data is imported without optimisations in the database (indexes/secondary keys), the time associated with handling the data and validating the extraction for correctness can become excessive. In some worst case scenarios it could take between 5 and 10 days before the archive is able to say whether or not the extraction is OK. It is common that the extraction

testing can fail for some reason, for example a referential integrity problem or duplicate values where it is prohibited. A new extraction is either generated from a database dump or the municipality creates a new extraction once the faults that caused it to fail are fixed and the testing process is again started.

This all points to the process domain as being an extremely important data quality domain for the handling and testing of incoming archive data packages. We have earlier stated that volume, time and technological obsolescence are inhibitors to data access. This is very true when looking at this issue from that archive data handler point of view. Production file formats from outdated software that are not easily converted to archive formats, lack of verifiability of converted documents are two such examples that municipality archives are facing on a daily basis. Some municipality archives are hesitant to handle electronic extractions as they simply do not know what kind of commitment they are setting themselves up for if they have to process, verify and maintain them.
Technology and time have also been good to the archive as disk space is relatively cheap. A 3TB disk can be purchased today for around 1 000,- NKR. A 512 GB high speed SSD is available for around 3 000,- NKR. Space and processing power has increased phenomenally the last number years so from a hardware cost perspective space and processing power is not expensive. But it is important to understand what an archive institution requires in terms of processing power and space to avoid generalisations. There is a major difference between storing (no:lagring) and preserving (no:bevaring) of Noark 4 data. The storage of files is not difficult at all, but preserving files with their integrity, authenticity and understandability is far more complicated

Data Quality and File Conversions

Data Quality and file conversion of files from a production format to an archive format (or a long term preservation format) is not a simple problem. The biggest problems facing data quality associated with file conversions and municipal records management is cost, volume, time, technological obsolescence and automation. These factors are also related to each other.

Types of files:

The number of various file types for the various files that are stored within a municipality Noark system is limited but at the same time challenging. Table 4.1 shows the number of various production filetypes in a number of Noark 4 systems and the long term preservation filetype.

Production Format	Archive Format	Count of documents in this format	Comment
BMP	RA-JPEG	100	
CSS	??	1	The contents of the CSS can be converted to TXT or PDF/A

DOC	RA-PDF	327 000	
DOCM	RA-PDF	12	
DOCX	RA-PDF	1 360	
DWF	RA-PDF	1	Possible for PDF/A export
DWG	RA-PDF	29	Possible for PDF/A export
DXF	RA-PDF	3	
EXE	N/A	5	An executable file should not be part of a Noark 4 extraction
GIF	RA-JPEG	1 227	
PDF	RA-PDF	342 000	PDF files scanned in or already converted to PDF/A
HTML/HTM	RA-PDF	19 000	HTML files were assumed to be single self contained files
JPEG/JPG	RA-JPEG	4 100	
LWP	RA-PDF		
MOV	MPEG2	1	
ODT	RA-PDF	25	
PPT	RA-PDF	280	
PPTS	RA-PDF	1	
PPTX	RA-PDF	30	
PNG	RA-JPEG	600	
RTF	RA-PDF	395	
TIF	RA-TIFF	493	
TIFF	RA-TIFF	2	
TXT	RA-TXT	42 461	
XLS	RA-PDF	958	
XLSM	RA-PDF	1	
XLSX	RA-PDF	110	
XPS	RA-PDF	1	
XML	RA-XML	4 410	Without DTD/XSD

ZIP	???	68	Manually processed to obtain contents

Table 4.1. *Count of various file types from a Noark 4 system*

Zip files have to be processed manually and luckily there was not many zip files in the database. The problem with a zip file or another container file format is that unless they are long term preservation compliant then we cannot use them. What happens if the zip file contains 10 other files in DOC format. The 10 files have to be converted to PDF/A and then new registrations will have to be made in the Noark 4 database for the new documents. There is software available that can process zip files but not that can add them to the extraction.

The volume of files within a municipality Noark system steadily increases over time and avoiding conversions to an archive format pushes a potentially expensive problem further down the road. The Noark system did at some stage have conversion of files from production format to archive format but not all PDF files were actually PDF/A compliant, something that adds another layer of complexity to the problem of creating extractions.

In terms of technological obsolescence the lack of available software to read and convert documents files can be problematic. Lotus Word Pro (LWP) and Wordperfect (WP) files are examples of files that can be difficult to convert to archive format. We note that over 200 000 document files were converted from a production format to PDF/A using PixEdit Pro v 7.11.22. This software was unable to handle Lotus Word Pro documents and some difficulties with Word documents (.doc) from 1998-1999, but we believe that this is a problem with file type detection not PixEdit. LibreOffice has builtin support for Lotus Word Pro and was used to handle the conversion of LWP files. However in some cases LibreOffice crashes with some LWP files. So even “modern” files can be difficult to convert.

Automation is the solution to handling volume. It is simply impossible from a cost perspective to manually convert many hundreds of thousands of documents from one format to another. Automation could be for example to use a tool like pixedit or to use a scripting language like php and MS Office/LibreOffice to convert documents.

A few years back the cost associated with the conversion of a single file from production to archive format for Word documents was estimated to be 2 NKR per document. This cost has reduced over the years. The total number of files in the collection of files that the report is based on is 850 000. At 2kr per file, the cost would have been 1 700 000 NKR just for file conversions. Having said that the collection of files includes over 340 000 PDF files where a good percentage of these are in PDF/A format.

None of the conversions had any check on content. Every converted document was only checked for PDF/A compliance. Such a check really is insufficient! An example that details why this is so is that a number of Lotus Word Pro documents were incorrectly identified as MS Office documents and were converted to PDF/A. The conversion software did not detect this error and simply produced a compliant PDF/A document filled with garbage. As an example of this is that a single page LWP file was turned into a 93 page valid PDF/A file full of binary symbols.

The closed nature of proprietary records management systems is also a potential hinder to tackling the problem. A relevant example of this closed nature is when dealing with production files that have been encrypted and the only way to automate an export of the files for conversion is to employ the original software provider to decrypt and extract the files. No software vendor should have that control over a municipalities files. There should be a greater focus of the use of APIs for these systems so it is easy for third parties to work on them.

Macros are probably the biggest worry when it comes handling documents in production formats. It is impossible to know what the macros do but one common we noticed was that they were being used for data entry. Often when processing excel files in Pixedit Pro, the presence of Macros would stop the conversion process with dialog boxes asking for information. From a cost perspective it is difficult to fix this.

We also note that on numerous occasions files that are listed in the database as being in PDF/A format are in fact only in PDF format. It is not certain why these files did not get converted to PDF/A. The main problem with this is the extra cost associated with detecting and converting these files to PDF/A. An additional question that should be asked is whether or not it is acceptable to convert the PDF file to PDF/A or if the original production format file has to be used i.e the original Word document that created the PDF file.

Pre processing of files

At some stage the Noark 4 system used had functionality that automatically inserted the municipality logo and other necessary information into a word document. This functionality has also been removed at some stage during the life cycle of the system and there are now a large number of MS Word documents that contain what appears to be pre-processing instructions. These are simple words and symbols like an exclamation mark. It is assumed that the contents of the file is correct but that some contextual information in the file is missing. In some regards the fact that this has taken place can to a certain degree cast doubt over the authenticity of files but given the associated Noark data (XML), we argue that these files should be seen as having maintained their authenticity.

File Detection

One of the biggest problems faced when creating the extraction was the detection of the production filetype. The Noark 4 system at hand used a proprietary method for naming file extensions and how it identified the file type. What we mean by that is that a *.doc* file extension in the database could be identified for example as a *.rcp* file . The application associated with the file is also stored in the database. However this information was not available for all files in the database and often incomplete and sometimes misleading. This fact meant that file type detection was more difficult than first anticipated. The approach we first used during the extraction was to rely on the database for this information. This is not a naive approach. One would assume that the data associated with files and file types would be correct and up to date. That assumption is based on the data quality with this type of data is high. The problem with file type identification also underpins the requirement for publicly available documentation. If an archive institution themselves had to process this database then the lack of documentation could be a severe hinder to access. We would not use the

same approach for file type detection again. However it is imperative that the systems do not make filetype detection difficult.

As the software to detect file types evolved a number of sources in the database were used. The first approach was to use the mime type associated with a file. This quickly showed itself to be an incomplete approach and a second source was located. The associated tool for opening the file was sometimes listed in the database. In some cases the file type for known files was already associated with the files. These include DOC, GIF, PDF. These file types were not tested, it was assumed that the database was correct.

So in all there were three sources of information to determine file type.

1. A given file type
2. a mimetype that was not always present
3. Associated tool for viewing document

The problem with 1. was that the field was not necessarily complete or adequately descriptive. When it comes to 2 and 3 usable values were not always present.

It was in many regards fitting for a Data Quality analysis to discover that bad data impeded the extraction of files from the database. In hindsight the approach should not rely on the data in the database and rely on automated detection with file type detection software. If this is undertaken the file type detection software has to be of such a high standard that it is capable of detecting all file types. The Linux *file* command is not powerful enough in this regard detecting doc/xls as MS office files not Word/Excel file and docx is detected as a zip file. A docx file is actually a zipped file containing xml files for content and any images in the document. So *file* is technically correct but not very useful in this regard.

At one stage when extracting files, five executable files were found as documents. We did not have time to analyse why these executable files were present, but the filename was outlook.exe and explorer.exe. It could be possible that somebody wanted to use the Noark 4 system to distribute this file or it could be related to some system install issue. There is an assumption that a Noark system contains documents, not executable files. An executable file poses a potential risk to an archive if it contains a virus. This really underpins the importance of quarantine and virus scanning as important parts of the process of handling incoming extractions at archive institutions. But the question could be raised on whether or not a system should we prohibit these files from being part of an extraction or even uploaded into the system.

Difference between File type detection and file type validation

When someone is new to the field of record management and archives, it is not always clear what the difference between detection and validation of files is or even that there is a requirement to be able to carry out detection and validation. An archive or municipality can be in a situation where they have many hundreds of thousands of files but lack the information regarding the type of file or the software that is used to open the file. This information is commonly identified by an extension and historically on windows systems file

type extensions were limited to three characters.
In this setting file type detection plays an important role in determining the software that created the file and can be further divided into two subtypes, 1) detecting the software that can handle the file and 2) where possible the version of software. For example a doc file could have been produced by many of the different versions of MS Word or even by third party applications. But it is not always clear which version of Word we should use or can we not just use the latest version as Microsoft Office to a really good degree has backward compatibility. We do note however that Word files from around 98/99 had problems being converted to PDF/A using the pixedit software. It could not handle these files and simply placed them in a failed queue. We did not have time to try and discover what was causing this but were able to script LibreOffice to undertake the conversions. The files could actually be LWP files detected incorrectly as DOC but as this project ended, we did not have time to get a final answer to the question.
From an extraction point of view, filetype detection is about detecting the software that can open a file for the purpose of creating an archive version. But it can also include detecting the version of the software if that is required. We had initially planned to use the DROID software for this but we had access to documentation about an early version of DROID that no longer was correct, so we actually wrote our own filetype detection tool using information from the Internet about signatures in files. It turns out that the DROID tool is sufficient for this job. This became nearly farcical as we have done a filetype detection process many times on the 850 000 files to determine the filetype and we have still not arrived at a final conclusion. This really should not be an issue and future Noark systems must be more robust with regards to filetype detection and archive versions.

File type validation on the other hand concerns itself with determining that a file of a given file type is actually a correct implementation of that file type. This is really important for archive formats. A PDF file is not necessarily a PDF/A file. It could be one of many PDF versions, for example PDF, PDF/E, PDF/UA, PDF/X, PDF/VT. It is essential that we can validate that a file that we believe is in PDF/A actually is a PDF/A file. We must be in a situation where we can independently validate any PDF/A file as conforming to the ISO standard describing PDF/A.

Data Quality Run for files

The closest we were able to achieve data quality in terms of file processing for an extraction is to test the files for archive format compliance. All documents were converted to an archive format but there were some issues that resulted in multiple attempts of conversion before an archive quality file could be produced.
The biggest problem we see with the current approach of automated converting of files to archive format is that it is extremely difficult to know that all the content in the converted file matches the content of the original file. Consider an MS Word file with an embedded MS EXCEL spreadsheet, this scenario is very difficult to automate content validation between production and archive format.

How to achieve Data Quality for file conversions

In 2010 we had supervised a masters project at DCU that resulted in the development of an open source framework for comparing converted files to their original and deriving statistical guarantees on the conversion process. This tool is called exactitude. The tool can be used

to check the conversion of multimedia files (images, audio and video) and can determine the similarity of original and converted file. Exactitude makes use of some external libraries that can compute similarity between multimedia files including the pHash library. This tool could be further developed to compare converted word processing files. Time constraints prevented us from using the tool to test the conversion of image files to jpeg but in the future we envisage using this tool.

Reflection

Within the community there are various ways this field approaches the lifecycle of records. Central to the term lifecycle we mean the three common phases records go through, records management, extraction and long term preservation. Some practitioners maintain that the lifecycle consist of three independent phases and the archive should not be involved during the records management phase. Some record managers also believe that the archive should not really concern themselves with what the municipality does during the records management phase. This difference in attitude is very evident depending on the background and the type of of archive you work in. Some practitioners maintain that we should not correct any data quality issues as we change the way history will perceive that extraction. By this we mean that a system that is not really capable of creating an extraction is now capable because we fix some data quality issues. That system will now have a different reputation compared to the archive only having an extraction with bad data. We agree with this observation and maintain that the archive should preserve both, the extraction with bad data and an extraction with good data. This allows both perspectives to be maintained.
When analysing structure and data quality issues it became to us clear that the life-cycle of electronic records should be viewed within a holistic approach, not just a series of interlinked phases. As a community we respect the varying opinions but we note that some practitioners have changed the minds when they look at their data in a database and they see their data differently when analysing problems inhibiting the creation of extractions. The consequence for the archive of this independent phase approach then is that the archive can and should not care what has happened but will maintain whatever records it gets. This approach adequately reflects a paper based practice but when the material is electronic this approach can be naive. The approach works fine if the records management phase has been handled properly and there has been a focus on data quality but blindly accepting extractions without fully understanding the records management process may be costly for the archive in the future.
We can develop automated tools that take care of a lot of the data quality measurements and that point out potential issues that should be manually inspected. Regular attempts to create an extraction from the various systems should be undertaken, perhaps as much as every six months. This could be undertaken on a backup copy where manual work is allowed to fix issues and test them again. To achieve this the software vendors will have to be involved and the cost issue may mean that it is not possible to achieve this.
As we delved deeper into the data quality research field it was difficult to maintain a presence in the archival theory field. But we ask ourselves to what degree can we include provenance, context, structure and fixity as issues that define data quality? The concept of a transaction has to be a core element of data quality. The whole point of a Noark 4 extraction is to capture provenance, context, structure and to a degree the fixity. If we take these 4 topics individually we see the following. Structure we can see in the archive structure, the fonds, series objects etc. There is a set of XML files that define this. With regards to context we see the documents set within a registration and casefile, which is integrated into the actual structure of (fonds, series etc). Provenance is probably a bit more difficult, especially in the system we worked with because we had to fix a number of issues with regards to people identified in the system. We can see provenance from a system perspective and from a records perspective in a Noark 4 extraction. Provenance from a system level is probably

over documented in a Noark 4 extraction, while provenance from a records perspective might be a little weak. If provenance is documenting history and ownership, the process information the records have been through are not visible. The access log and change logs are lacking. Fixity is not a central element of Noark 4 (it is however central in Noark 5 extractions) with perhaps NOARKIH.XML being the most obvious fixity element. It can be argued that as long as the archive structure is intact and the casefile, registration and documents are present and correct the concept of a transaction is correctly documented. In this regards many tables in a Noark 4 extraction serve little purpose and in fact end up complicating the extraction process as some vendors implement the standard loosely.

We also believe it is also important to never touch the original Noark 4 database, but always work on a copy. Washing data and fixing problems on the original should not be undertaken as you may inadvertently create problems that later can not be fixed.

Finally we make a point that from a data quality perspective when considering data from various municipalities with data various Noark 4 production systems, it is most probable that the data from two various Noark 4 systems will not be technically comparable. By comparable we mean that tables will have different names and that identifiers can follow different conventions. This is a syntactic issue that naturally results in semantic issues. We also observed various use of identifiers in the same Noark 4 system across municipalities. By this we mean the format of for example case file identifier. These case file identifiers/ document identifiers are defined with various lengths. In 5 databases from the same system the same identifier format was used for case files but for one of the databases an extra character was used for this identifier. This was often just an extra 0, but we have no understanding of why this situation arose. This anomaly can be an inhibitor for automated processing of data within extractions.

5. DQ Comparison (Deliverable 4)

In many ways it is expected that this report should be very short. There really should be no difference between the data in the system and the data in the extraction. When developing the project it was anticipated that there *could* be something left behind and we should at least look at the database to make sure all data has been extracted. We could nearly anticipate that the further the deviation from strict adherence to the Noark 4 standard the greater the chance the difference will be. If the system under investigation had followed a strict implementation of the standard then this report should really just note that everything was extracted and the data quality is the same between the two instance of the data. However when considering the user then data quality becomes a lot more complex and this really is a classic issue with regards to the migration strategy. The data goes through an existential change in functionality and use where the user is in many ways eradicated through the process of creating the extraction. We should clarify that the user of the data in production format is eradicated while a new user, the one in the archive gains visibility.
The vendor did not follow a strict interpretation of the standard and this means that we are left fixing issues to make the data quality acceptable from an archive perspective. The more we change the data the greater the difference in data quality in the two copies of the data. Not changing the data in the production system allows it to preserve its data quality for the end users i.e the case handler and records manager. But the archive can not work with data in this realm. It needs the data in a neutral format that can be preserved over a very long period of time.
In many regards this scenario is the accuracy data quality dimension and we should be able to measure the difference between the source database and extraction. As earlier we can denote this **v** and **v'.** When the Noark 4 database structure is interpreted and implemented loosely the difference between **v** and **v'** can be an important issue. There are two main ways that **v** and **v'** can differ. The first is that a lot of the data has to be washed so that an extraction can actually be undertaken and the second is that there is data in the database (v) that has not made it to the extraction. If the difference between v and v' is significant enough then the archive will be required to store various versions of the data. Potentially four different versions of the data have to be stored.

1. Valid but potentially incorrect
2. Invalid but Correct
3. Original database dump
4. Database in a neutral format

The first example is a valid extraction in the sense that it can be validated with a tool like arkn4 or URD but it is incorrect in the sense that data has been changed. By *potentially incorrect* we mean that the original content in the Noark 4 database structure has changed significantly and we might not be sure about the process that changed them.
The second version that is invalid but correct implies that the data was extracted directly to a version of a Noark 4 extraction without changing anything. Referential integrity problems and missing data can render the extraction invalid by validation tools but the extraction itself is a correct reflection of the contents of the Noark 4 database.

The archive institution will in this case probably wish to maintain a copy of the original database dump for as long as it is possible to access the database dump. Questions about the integrity of the database make this necessary.
At some stage it is likely that the database dump will become unreadable and in this case a copy of the database in a neutral format like XML, ADDML will be necessary to ensure access to the data. The SIARD tool could also be used here.
A reason why the database dump file must be stored in addition to the database contents in a neutral format is necessary is for example the case where the files are stored in the database using some non-universal datatype as discussed in Deliverable 3. In this and other cases where binary information is stored it may be necessary to store the original database. This approach may seem like overkill for a Noark system but in a case where there is no documentation of the database the two interpretations (valid and invalid) of Noark will be very useful in understanding the structure of how the original database implemented the Noark 4 standard. This approach also underpins the need for standards like Noark to require strict implementation of extraction functionally.

Another observation is that the municipalities tested a module that registered which politicians came to meeting and if they were owed any dues for attending. This module was only used for a short period of time (max over a few months) before it was discontinued. So there is some data in the tables. The temporary use of this module really shows how data quality can be a problematic issue. If this database dump had been the only source available in the future and analysts had to extract the data, the most natural approach would be to try and convert it to a Noark 4 extraction. The reason for this is that it would guide the search and understanding of the database and records contained within. The researchers would have found some data that showed the politician module was in place but that only a few months worth of data was there. What would this mean with regards to trust of the database as a source? If there is no record of the module being installed and subsequently discontinued we can and should ask the question, what other data is missing. We were also unable to create an extraction of the data from this module in the required Noark 4 format as some obligatory fields were missing in the production system.

An approach to understanding Noark data better is to do some statistical analysis on the data. A lot of interesting and useful questions can be asked in order to better understand the data. An example of this is the number of cases the municipality handles over any given time period. If this can be measured statistically then we might be able detect periods of missing data and this becomes an additional data quality dimension that an archive checks for.

One approach to handling missing or incomplete data is that the code that creates NOARKIH.XML file should be made from the source database and not made when processing the XML files. Recall that NOARKIH.XML contains the name of each table and the number of records in the file. While we do not have knowledge about how the vendors develop their code to extract data from the database, it would not be hard to imagine that NOARKIH.XML could be a result of running counters being incremented each time a record is processed while an extraction is being created.

Difference in DQ Before and after

How different is the data quality in the Noark system compared to the data in the extraction. If you ask the municipality, the users and records manager they will probably find the data in the system a lot more useful than data in the extraction. The first reason for this has to do with access. It is unlikely that the data in an extraction will find its way back to the production system and accessing data in tool like URD or directly in a database will not be a natural way for a case handler to work. However the data the case handlers regularly work with will probably be just as correct in the extraction as it is in the Noark system. The case handlers will mainly be concerned with cases, registry entries and documents. These have pretty good metadata and will be intact. Many of the data quality issues we have seen have not been related to these kinds of data. The biggest issue probably was the files but each file is converted to an archive format and understandable by the user.

So if we ignore the user interface and concentrate on the data from a users point of view the core elements of an extraction will probably have a high degree of similarity between **v** and **v**'. As we focus the comparison of v and v' beyond these core data object we will expect to see the difference between v and v' increase. Anything that is not central to a case file or archive structure will probably see increased dissimilarity between **v** and **v'**. We see this in the extraction we worked with, the core objects for preservation are there with little DQ issues and the DQ issues discussed in Deliverable 3 really relate to surrounding information.

If we raise our data awareness beyond what is defined within the extraction and look at what additional data is in **v** that is not in **v'** we can focus on both Noark data and other data. The example earlier about the data in the Noark system about political administration where a functional module was used for a short period of time and then discontinued is a good example of how **v** and **v**' naturally will differ for Noark data (providing we did not include the data in this temporary module). Other data included in the database that will not find its way to the extraction are:

- System information that details which updates have been run and when
- Information about user searches
- Access logs

There is no simple answer to this question and if we were to delve deeper into this issue we would have to study the data in the system and this could lead to concerns with regards to privacy. However a lot of the remaining data is not really personal data, but we are not in a position to make that call.

So it is difficult to answer the question of how different the data quality is when comparing **v** and **v'** but we can theorise that the closer the system is to the standard, the smaller the difference. The further the system is from the standard, the greater the *potential* for a larger difference. However this fact remains unproved

6. Data Quality Workshop

The project was aimed at the municipality archive sector and the project members believed we should try and present the results to the archive community to foster further discussion and interest.

Workshop

The two most important contributions to the project from the workshop were the identification of two roles that were overlooked and a desire to differentiate the identification of data quality from the identification of how an extraction deviates from the Noark 4 standard.

Deviation from the standard is a clear data quality issue, but to a certain degree within records management and archive data quality will not and should not be fixed. Fixing bad data in a Noark extraction can be seen as changing history with regards to the original content. We agree with this point and point that we argue that when the data quality is so bad that a validated and correct extraction can not be produced then there is a requirement to store different versions of the database so that history is not changed.

With regards to the roles the workshop attendees identified the *application developer/ software vendor* and the *entity creating the extraction* as two roles that should have been identified. Roles identification was never an aim of the project but rather something we discovered through discussion and analysis but we agree with the workshop attendees that these two roles also are important and any future work should include them. Interestingly the role of creating the extraction was the role we ourselve played when we wrote the code to extract data to an XML format. We became blind to our own role.

If an archive is only to maintain the original extraction without fixing bad data quality they are putting themselves in a situation for potentially increased preservation costs in the future. These costs will be related to search and access. This factor may in the future lead to a “data delayed is data denied” as it might be difficult to locate records in a timely fashion.

To a certain degree the Norwegian approach to long term preservation follows the traditional cardboard box approach where documents etc are stored in boxes and transferred to the archive. Search based on content on material in a box is not something one traditionally has been able to undertake and the same thinking is being brought into modern electronic archival practice.

One of the participants made an interesting point arguing that the lack of data quality during the records management results in a kind of technical debt that the municipality accumulates. Technical debt is a phrase that has come from work within the Norwegian Computer Society that tries to define the cost of not dealing with a backlog of technical issues. He pointed out that this approach could be useful when trying to make the municipalities understand the cost of not undertaking extractions in a timely manner. Technical debt and archives is an interesting future work aspect.

A lot of the delegates, especially the ones that work with extractions easily recognised the various issues that were brought up. Data Quality is something that the municipality archive

community already work with at various levels. In many ways this project is the first step at merging the formal research area known as Data Quality with Norwegian Archive tradition. Both these fields are compatible with each other and overlap and share commonalities.

Questionnaire and Results

We undertook an online questionnaire after the workshop to see what the attendees think about the project. We asked six straight forward questions. Without the project members there were about 16 attendees. Of the 16 attendees, 11 answered the questionnaire. During the workshop we presented project and the results of our analysis. There was a lot of dialog and sharing of opinion during the workshop and we got further insight into this area from a technical, archival and historical perspective. The attendees were asked 6 questions as detailed below. When we use the word “results”, we mean the data quality analysis and approach undertaken in this project with regards to Noark 4. The results presented here act as an evaluation of the project by the intended users of the results.

Question 1:

This question asked if the attendees find the results of the project relevant from a records management perspective. Most attendees found the results relevant or very relevant to records management. This is shown in Figure 6.1.

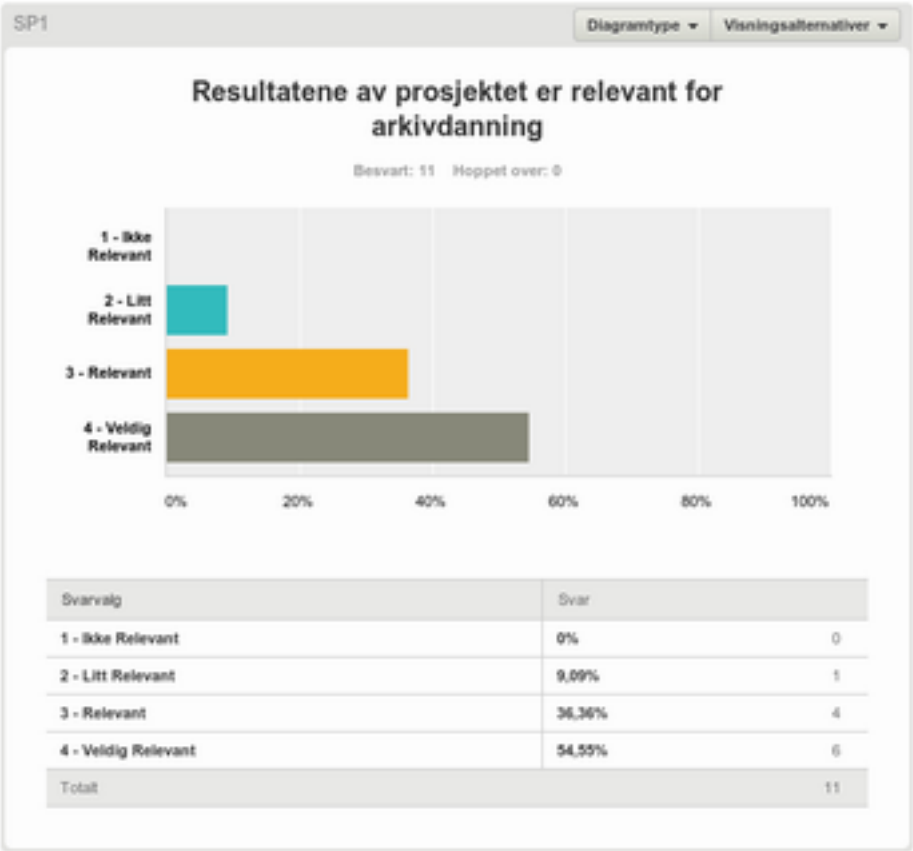

Figure 6.1: *The results of the project are relevant from a records management perspective*

Question 2:

This question asked if the attendees find the results of the project relevant from an archival perspective. This is shown in Figure 6.2. All attendees found the results relevant or very relevant to records management. Interestingly one attendee differs in opinion on relevancy between records management and archival perspective on data quality. It is difficult to interpret this discrepancy.

Figure 6.2: *The results of the project are relevant from an archival perspective*

Question 3:

This question asked if the attendees find the results of the project are relevant to their daily work. Again Most attendees found the results relevant or very relevant to their daily work. Not all attendees work with extractions so for some this work could be a little abstract. The results are shown in Figure 6.3.

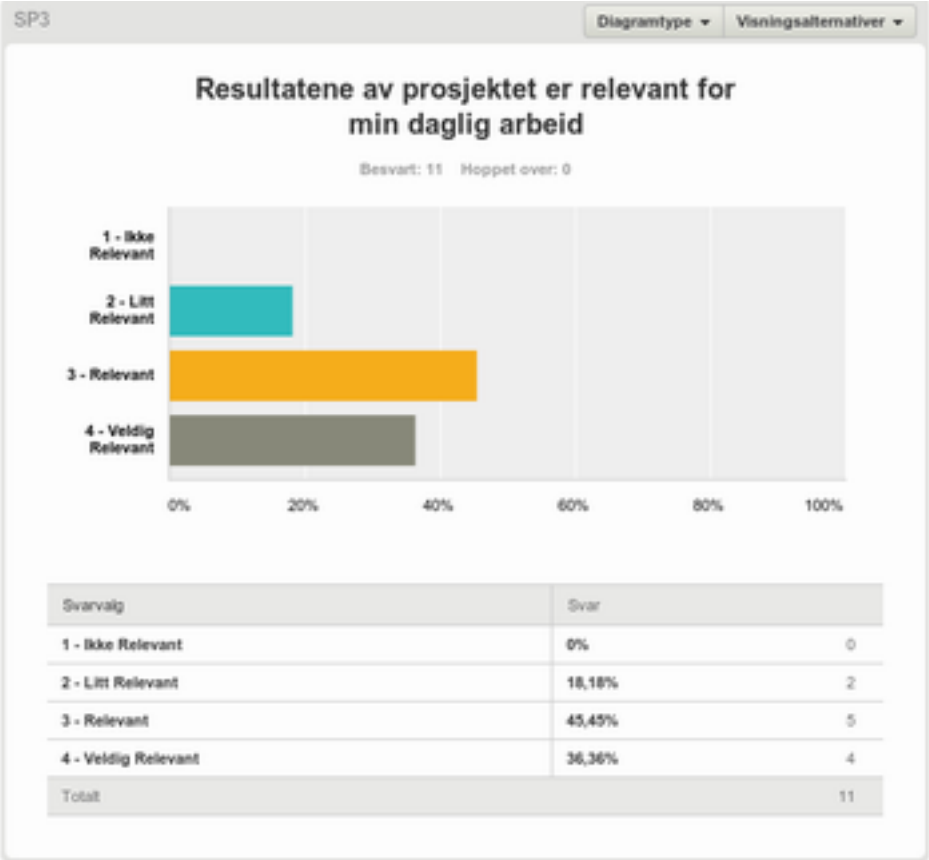

Figure 6.3: *The results of the project are relevant to my daily work*

Question 4:

This question asked if the attendees find the results of the project are relevant to how they will work in the future. Just over half the attendees say it will have an impact on how they work in future. 4 of 11 attendees state it will have a little relevance for future work. The project may come across as slightly academic and therefore it can be difficult to see how the results are used. The results are shown in Figure 6.4.

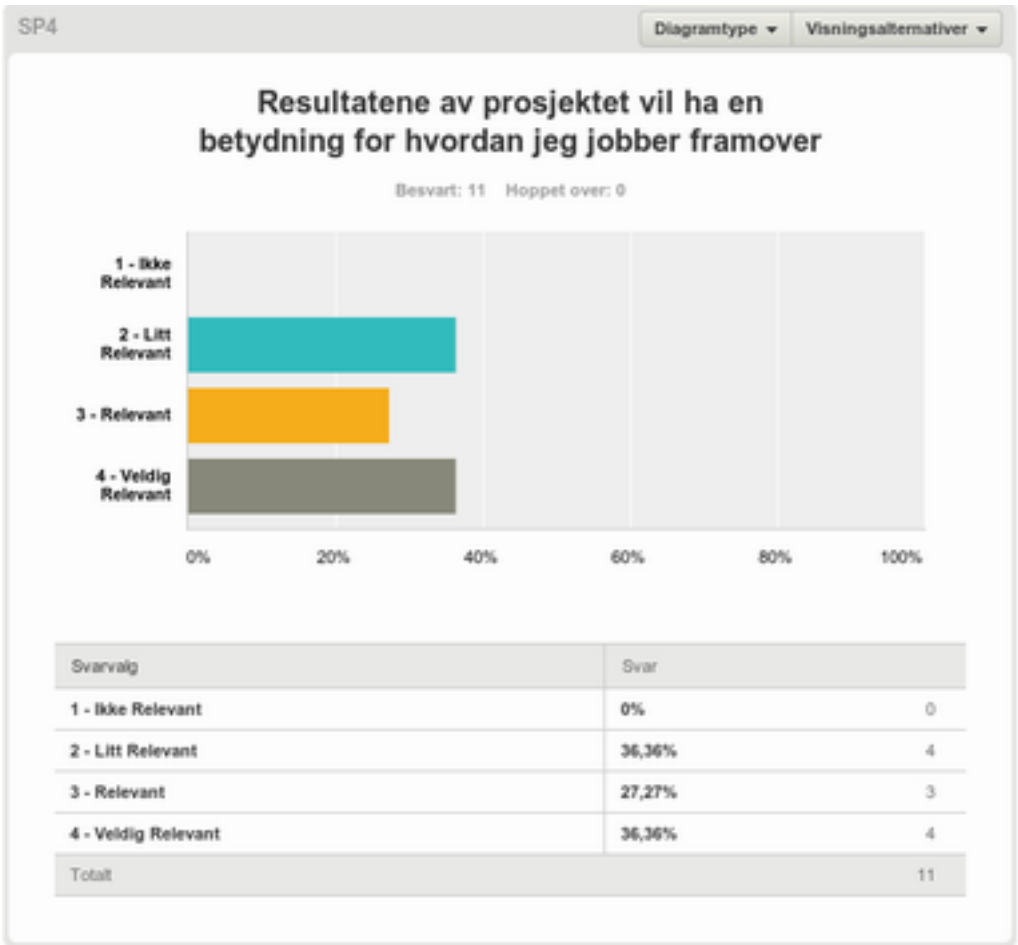

Figure 6.4: *The results of the project will have an impact on how I work in the future*

Question 5:

This question asked if the attendees believe there should be a greater focus on data quality during the records management phase. 8 All attendees either really agree or agree with this statement. We make no reference here about how this could or should be achieved. The results are shown in Figure 6.5.

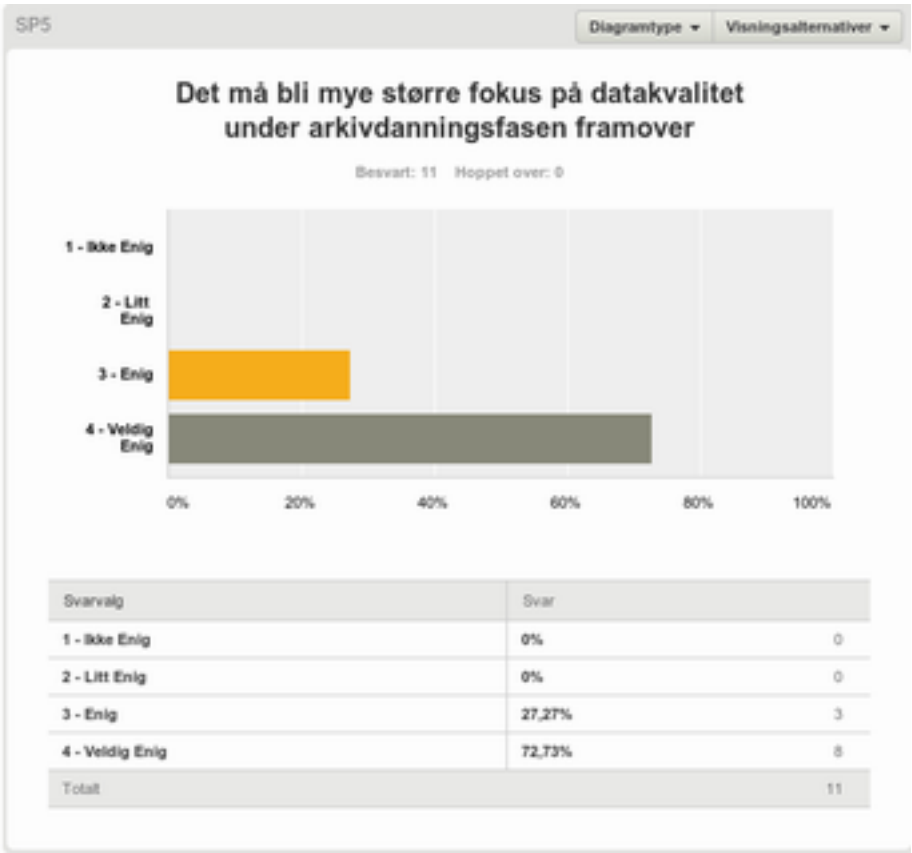

Figure 6.5: *There should be a greater focus on DQ during the records management phase*

Question 6:

This question asked if the attendees believe the project captured the essence of what data quality is for Records Management/Archives. The results of this are shown in Figure 6.6. All attendees agreed to some degree that this was true but only 3 attendees really agree with this. From the discussions during the day it is clear that the attendees all had different interpretations of what data quality is. The topic of data quality being about a deviation from the standard stands out in our minds as one of the reasons for not scoring higher here. Also some of the attendees do not work electronic extractions on a daily basis.

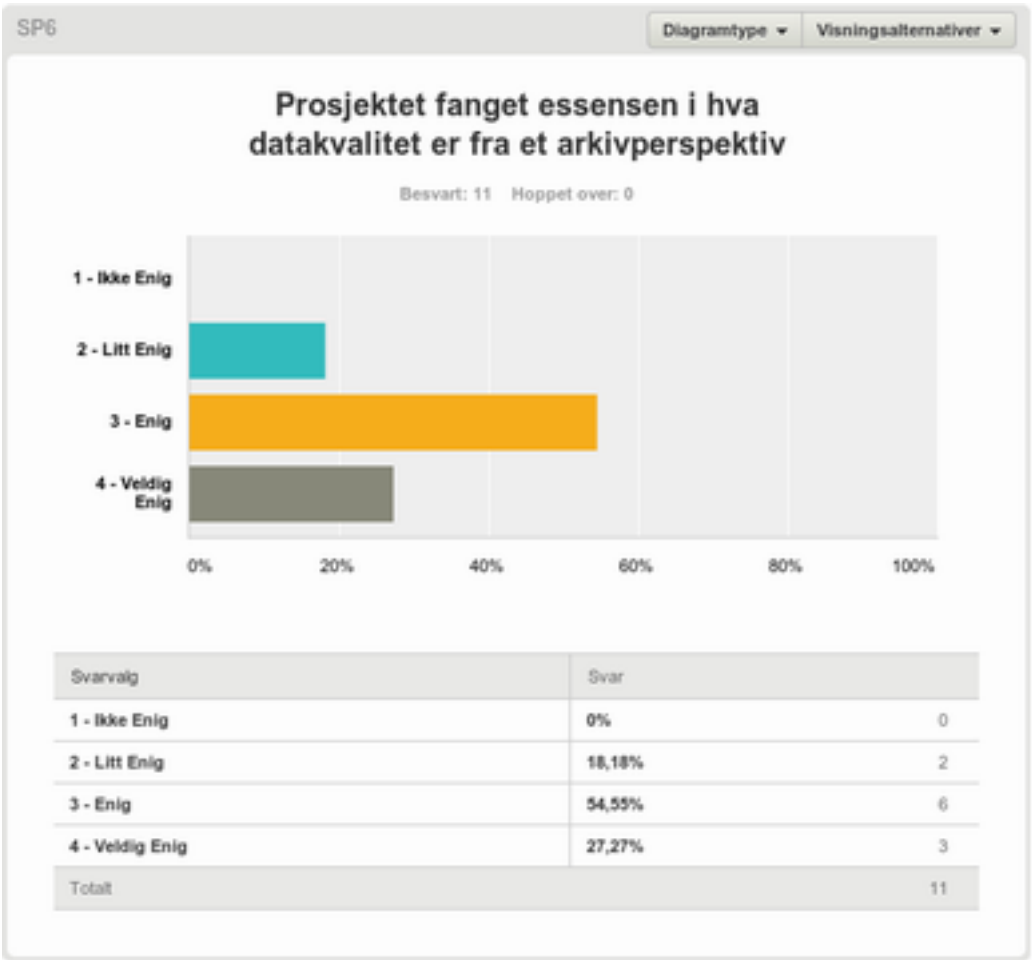

Figure 6.6: *The project captured the essence of what DQ is for Records Management/ Archives*

Summary

We believe the results of this questionnaire show that the project has delivered a report that shows its relevance to municipality archival community. We used the workshop to present our work and engage in dialog. We think that it can be hard to see beyond syntactic and semantic data quality and the pragmatic approach to data quality might come across as abstract. It is clear that data quality is an issue everyone that works with electronic records is concerned about but most institutions have varying approaches to what data quality means and how it should be part of their daily work.

Chapter 7. Project Findings

The project started with the development of a definition of what we believe data quality is. The initial definition was "***data quality is the degree to which data in a system reflects the real world scenario the data represents***". Initially this definition seemed to correctly capture the extraction scenario as it could equally apply to both the records management phase as well as the long term preservation phase. However the field of data quality has a very strong focus that includes the definition of a user as what is the point of collecting data unless it is usable by a person for a purpose. We saw, at the beginning, that archival theory and the use of the migration strategy seemed to be at odds with the data quality field. It can be argued that the Norwegian model of migration as exercised through the archives laws and regulations deliberately attempts to remove the user. The focus is on documents and metadata to preserve integrity, authenticity, provenance, context and structure. Discussions led us to believe us that even though the migration strategy does not explicitly include a defined user, there is neither an exclusion of a defined user. An extraction from a Noark (both 4 and 5) system is very much a system centric approach to preserving data where the data within the context of a system is preserved for future generations.

In the end the definition was changed slightly to include the definition of a user and became "***data quality is the degree to which data in a system reflects the real world scenario the data represents and is usable***". This definition also captured all phases of the life cycle of Noark records but at the same time made an important distinction that there has to be a user. We do not state who the user is as we do not know who the user will be in 100 years time.

As a result of the inclusion of a user in the definition of data quality, the need to look at data quality from the user perspective became more important. Deliverable 2 evolved from being a description of data quality dimensions relevant to Noark 4 to a description of data quality dimensions relevant to Noark 4 from the *perspective of various users*. This deviation resulted in the deliverable being much longer than what we initially expected but it really set the stage for interesting discussions where we tried to develop examples of the how the various users could perceive data quality. From a teaching perspective, the topic of data quality and Noark is often abstract and it is difficult to delve into the various dimensions within a records management and archival perspective. This analysis left us with more questions than answers and we believe this topic should be explored in more detail by the community. It was also at this stage the project moved from having a syntactic and partially semantic approach to data quality to having a more pragmatic approach. Data Quality was no longer just about measuring something in a database, it was about reflection and deeper understanding.

Within the scope of the project we did not carry out research on data. We would not have been allowed to work with personal data from a research perspective. Instead we aligned ourselves with a project that was developing code to create an extraction from a Noark 4 system that was rumoured to have known problems with regards to extractions. We also had attended various meetings with practitioners representing records management and long term preservation.

The combination of the above formed the basis of Deliverable 3. In Deliverable 4 we asked the question, whether or not there is a difference in data quality in the database as opposed to the

extraction. We could not base the answer to the question on a Noark 4 database, rather we looked at what the extraction code had done and not done. Then, when we included the concept of a user, the answer was far from obvious.
We held a one-day workshop at the end of the project to present the results to the community who would most likely find the results of the project useful. Nearly as anticipated, there simply was not time to discuss all aspects of data quality within a Noark context. The community identified and presented additional roles that should be part of this discussion and different perspectives on how they view data quality and we again were left with more questions than answers.

As we approach the end of this project, there are a number loose ends left. There are things that we wished we could have explored in greater detail but there simply was not enough time, there are new questions that get raised and there are new directions we can take. The rest of this chapter explores these issues and set the stage for future work.

Ravn and Høedholt have an interesting diagram that they use to describe Data Quality. This is shown in Figure 7.1.

Figure 7.1: *Data Quality Dimensions*

They define the following dimensions as being important:

- Completeness
- Validity
- Accuracy
- Consistency
- Integrity
- Timeliness

We had already covered these dimensions through the analysis of Wangs work and data quality but Ravn and Høedholts dimensions are presented more simplistically and probably capture in a nice way some of the important dimensions from a Noark perspective. These are really only a subset of what data quality is from a records management and archival perspective, but they make a good starting point.

There are many issues that lead to bad data quality and as with previous research within the data quality field we saw that “bad” system design was a definite factor. By “bad” we mean a loose interpretation of the standard, the lack of use of primary keys, referential and so on. Yes, some issues were related to the end user but the system did seem to cause the most problems. We are not sure of this fact as we did not have a chance to work directly with data so we did not get a chance to explore the user side of things. The pragmatic approach leads us to believe that this is probably a correct observation. It would be very interesting to see how the data quality is in a system that has a strict interpretation of the standard.

Throughout the discussions and analysis two factors that were the biggest problem to dealing with data quality issues became apparent. They were simply *time* and *volume*. These two factors really are the biggest hinder in fixing data quality but there is not really a need for them to be an issue. We really need tools to measure data quality continually and fix issues as they present themselves, rather than waiting to accumulate records over a number of years and not be able to fix the issues due to cost.

We learnt that the field has no unanimous view on data quality and some practitioners want the extraction to be a correct representation of the actual records management process. Some practitioners are very adamant that we do not fix data quality as by doing so we are changing history. We think there was consensus that maintaining the original extraction with bad data alongside with a version that has good data quality is an acceptable middle ground. But this will potentially increase the cost associated with long term preservation for the municipal archives. Mitigating costs for a municipal archive is an important issue and today's situation with bad data quality does not give any encouragement that we are heading in the right direction. The General Auditors report on the state municipal records addresses many issues and it is unclear what practical steps are in place to mitigate the problem. Technical competency is a clear issue but for the municipal archives to work at this level we will require access to system documentation. This kind of documentation may be considered a business secret from the vendors point of view but there is a clear and urgent need for archival institutions to have access to such information. Are the municipal archives capable at working at this level. The answer to that is Yes. There is a lot of knowledge within this community and KDRS are doing a good job make sure it get spread. However the various archives have various approaches with combinations of proprietary and open source tools being used. There does not seem to be a consensus around the way forward but it is heading in the right direction.

From the records management side of things there can also be misunderstanding by the IT role in the municipality. On a few occasions the IT representatives were under the impression that a municipality could deposit their electronic records with the National Archive and the municipality had no role to play in terms of the long term preservation of their own records. This notion is wrong. The municipalities are themselves responsible for the long term preservation of their records and often use the services of a non-profit inter municipality archive.

We have also seen how municipal archives can come under pressure to formally accept an undocumented extraction consisting of a database dump with associated files. The associated files are encrypted and no information is provided on how to decrypt them. There seems to be an expectation that the municipal archives can undertake miracles in this kind of scenario. Another disturbing observation is that the municipality archive can be put under pressure to say that they are now in a position to undertake the long term preservation of an extraction even though the municipality has not provided a complete extraction. We saw examples where the documents are referenced on network drives and perhaps the most ugly case when the entire database provided by the municipality referenced the actual database file on a network location. There does appear to be a disconnect between the IT department in a municipality and the data they have a legal obligation to preserve.

Another issue we came across is perhaps a myth, we were often told there that there are no long term preservation problem with the Noark systems the municipalities are using. It is the specialised systems (fagsystem) that are problematic. This notion is misleading and in some cases downright incorrect. The ability to create extractions from Noark 4 systems is something that sometimes is so prohibitively expensive that a municipality can not afford to create them. Instead the problem is pushed forward in time.

There is also a common misconception that technology has made electronic records management far easier than paper based records. It is not uncommon to hear this from decision makers at municipality level in Norway. This notion, that electronic records is easier to preserve than paper records, is so wrong. Electronic records management is far more scalable and offers the potential of easier collaboration but a lot of the issues we have identified are limited to electronic records and do not manifest themselves for paper based records.

The case was made a few times that when it comes to paper based records, post-it notes that were put on top of the case file were consistently left on the casefile even when the casefile was archived. A case handler might make a comment like “Client is difficult and unwilling to cooperate”. These kinds of comments give valuable insight into how the case was handled and can even be used to show that the case handler showed prejudice. This type of information falls so much quicker away in the electronic records management systems. In this regards the electronic records systems will have less quality than paper based records as the underlying medium does not support an easy way to handle these kind of comments.

We also note an interesting situation in the United States where there is an Information Quality Act[1] to ensure a minimum level of data quality. A similar approach could be undertaken in Norway.

Data Quality is a journey that everyone who works with electronic archive material takes and is something that often means different things to different people within this community. It is clear that this topic should become a focus at an organisational level for all archive institutions and it should work its way back to the municipalities.

After seeing the issues raised in the report it is natural to ask how the issues can be fixed. Time and volume are the biggest contributors to the cost factor when working with extractions but also when aiming to maximise data quality. Any key approach to increasing data quality for municipalities resides in eliminating time and volume as data quality inhibitors. Data Quality analysis tools is probably the most cost effective way of achieving this. These tools should perhaps be developed for the individual roles as described in Chapter 3.
It can also be worthwhile to show the record managers their own data in the database and discuss the migration strategy for long term preservation and explain how decisions made during the records management phase have potential implications for long term preservation.

It is also important when working with data quality to not point fingers. Data Quality should be more about raising awareness that saying proving that a particular municipality has bad data. However the importance of this issue needs to be raised.

Data Quality really complements records management and long term preservation, but the next step here should be to align Data Quality with the classical archive dimensions of provenance, context, structure and fixity.

As the project closes we really are left with many more questions than answers. We need to further our understanding of data quality from the various roles identified in Deliverable 2. Are there any micro/macro effects that are apparent in data quality of Noark systems. Time and volume are clearly the biggest inhibitor but what other factors should we consider. Can we quantify data as a single decimal number between 0 and 1? Should interoperability be a data quality domain?

With regards to future research there is certainly scope for another project that can naturally pick up from where this project ends. It may make sense in the future to include practioners at the state level and include Statsarkivet or Riksarkivet. The data they have will certainly by different to the data the municipalities have. To a certain degree this kind of project should again be funded by Kulturrådet or Forskningsrådet as we have not reached the point where the results are readily integrated back to the community. At some stage a DQ project like this should effect a change in the way the practitioners work and any future work should also include a strategy in how to get the results used in practice. There is also scope for this work to be funded at Ph.D.

[1] http://www.whitehouse.gov/omb/fedreg_final_information_quality_guidelines

level funded by Forskningsrådet but it is unclear where this work fits in. It could be under the verdikt program or a social sciences program.

Future work should include specialised systems and through the discussions we have undertaken we note that the migration strategy in Norway really is based on system approach. We should move away from this system centric focus to an object centric focus. An object centric focus will allow the archives greater flexibility to preserve and access data from all the various systems where objects of interest can be defined and retrieved.

http://xkcd.com/1179/

8. Cost Report:

Budget as per project application

Hva	**Beskrivelse**	**Sum**
Personalkostnader	HiOA	kr 304 335,00
	KDRS	kr 95 000,00
	IKA Kongsberg	kr 87 500,00
Tjenester kjøpt	Helfert Markus	kr 81 250,00
Andre utgifter	Reise	kr 98 000,00
Totale utgifter		kr 666 085,00

Tabell 1. Oversikt over forventet bruk av midler slik det var spesifisert i prosjektsøknaden

Oversikt over midler som er bokført per 28.06.2013:

Hva	**Beskrivelse**	**Sum**
Personalkostnader	HiOA	kr 304 335,00
	KDRS	kr 95 000,00
	IKA Kongsberg	kr 87 500,00
Tjenester kjøpt	Helfert Markus	kr 81 250,00
Andre utgifter	Reise/Workshop	kr 82 840,00
Totale utgifter		kr 650 925,00

Tabell 2.Oversikt over midler slik de var brukt i prosjektet

As per 28.06.2013 the final accounts have not been prepared and we are still awaiting some invoices. The final accounting figures will be sent to kulturrådet as soon as they are available.