Getnet Lemma Tefera

Implementing an XML Object Identification System on an archive data

Supervisor: Thomas SØDRING

Oslo University College

FACULTY OF JOURNALISM, LIBRARY AND INFORMATION SCIENCE

Master thesis

International Master in Digital Library Learning

2011

Dedication

To my beloved Parents, for all the love, support and encouragement you offered me!

Acknowledgement

I am very grateful to my Supervisor Thomas Sødring, without your support and guidance it would have been difficult, Thank You!

Special thanks to all the DILL Community in Oslo University College, Tallinn University and Parma University, and all the visiting professors, I am grateful for all the Knowledge you have shared, the support and guidance in these two years period.

I would like to thank The European Commission for the funding of my study through The Erasmus Mundus Programme, without which I couldn't imagine this study.

I would also like to thank my Dear DILL Classmates; you all have been great brothers and sisters, without your friendship it would have been difficult to accomplish this study.

I would like to thank my Parents for their sacrifice to educate me and to their great parenthood, I am Very grateful!

My grateful thanks also go to all my families, I wouldn't imagine today without your love, support and encouragement!

Abstract

Despite the existence of various techniques and tools at early stage, the data quality problem was not given the attention it deserves, until recent time,1990s the data quality was restricted to certain sectors, but later following the exposition of the huge losses due to data quality related problems different works has been seen. A few scholars have been involved in exposing the data quality problem and also finding solutions; among the initiatives to study the data quality problem systematically was the total data quality management methodology.

The archiving sector is not a different from the above case, in the process of archiving or long term preservation unless the data preserved is accurate and authentic its use would be of little value.

This paper is the study of how to ensure the accuracy of digital archives data and it presents a data quality approach called an object identification technique as a way of ensuring that an archive data is accurate. Most of the research undertakings have been focusing on relational data, but with the increasing popularity and importance of the XML data, there is a concern for developing data quality tools and methodologies which suit the XML data need. Based on this fact the object identification technique on this study focused on an XML data.

The research used the Noark data as a case study and developed a prototype of an object identification technique. The prototyped object identification technique has shown a good result upon a test on sample Noark representative data.

This study is of significant in taking the initiative to create the awareness on data quality issues in the case of an archive.

Keywords: Data quality, Object identification, Noark, Archive, XML

Table of Contents

Abbreviations

TRAC	Trust worthy Repositories Audit and Check
XML	Extensible Markup Language
NOARK	Norwegian record keeping System
ICT	Information communication technology
OCLC	Center for Online Computer Library Center
CRL	Center of Research Libraries
SIP	Submission Information Package
DQ	Data Quality
DoD	Department of Defense
ETL	Extract Transfer Load
EIA	Energy Information Administration
TDQM	Total Data Quality Management

List of Figures

List of Tables

Chapter-1

1.1 Chapter Introduction

This chapter states the motivation behind the research, followed by a brief explanation of some of the concepts that lead to the problem statement and the research questions. The aims and objectives of the research are introduced and the scope of the research, target audience, significance and possible limitations of the research are discussed.

1.2 Motivation

Electronic data play a crucial role in the information and communication technology (ICT) society; they are managed by business and governmental applications, by various applications on the web, and are fundamental in all relationships between governments, businesses, and citizens. Due to the nature of information technology electronic data can be easily be shared by many people and systems; the implied or observed "quality" of such data and its related effects on every kind of activity of the ICT society are becoming more and more critical.

It is widely known that problems related to data quality cost different organizations, businesses, corporations, non-profit, significant amounts of money in addition to other losses which cannot easily changed into dollar amounts. The data quality problem costs U.S. businesses more than 600 billion dollars a year (Eckerson, 2002). The widely known ""Year 2000 problem", led to modifications of software applications and databases that used a two-digit field to represent years, has been a data quality problem and "further the costs to modify such software applications and databases have been estimated to be around 1.5 trillion US dollars"(Batini and Scannapieca,2006). A more grim and series example that vividly illustrates the consequence of bad data quality is the explosion of the space shuttle challenger that was in part attributed to data quality problems (Fisher and Kingma, 2001), as cited (Batini and Scannapieca, 2006).

The above examples illustrate the data quality issue, and not to help solve the problems is the fact that organizations are not aware of the size of the problem. According to the data warehousing institute data quality report by Eckerson (2002) there is a gap in perception and reality regarding data quality in most organizations. As an information specialist the researcher has got interesting to contribute to the field by engaging on activities which can create a better awareness of problems, research and solutions to dataquality.

1.3 Introduction

According to the 2010 “the digital universe” report, the amount of digital information generated is estimated to be in excess of 1.2 million petabytes, and in 2007, for the first time in the history of mankind it is estimated that the amount of digital information created exceeds the amount of available storage (Digital Universe, 2010). This dramatic increase has brought a major concern on how to deal with information: How will we find the information we need when we need it? How will we know what information we need to keep, and how will we keep it? How will we follow the growing number of government and industry rules about retaining records, tracking transactions, and ensuring information privacy? How will we protect the information we need to protect? Among the implications is the need for good information management, ways to structure unstructured data and more compliance tools; organizations are expected to implement information life-cycle management as an enterprise wide information management strategy. Data once created, will serve a purpose and at some point its usefulness and need to exist will come to an end, as it may further updated or actively read, this can create a problem in databases or data warehouses holding such data with a related consequence on performance. For some organizations there may be a need for digital archiving that is transferring the less active data to another repository.

What constitutes the business case for archiving? Organizations need to archive their data for different reasons, among them are: legal actions, if the organization fail to discover its data it may lose legal cases, it helps savings in storage technology, administrative costs, backup profile and performance and deferred system upgrades (Agosta ,2008). Digital archives and digital material in general can be viewed as being highly fragile; their existence depends on technologies that are continuously and rapidly changing. Technological change ensures that within short period of time both the media and format of older digital materials can become unusable (Flecker, 2003). ‘’Digital preservation is the set of activities required to make sure digital objects can be located, rendered, used and understood in the future’’ (Digital Preservation Europe, 2006).

The preservation of digital materials alone could not guarantee the quality of preserved data:

> Even if we could ensure the preservation of electronic entities and overcome media fragility and technological obsolescence, preserved materials would be of little value unless we can be sure that they are 1) accurate, that is, precise and free from error or distortions, and 2) authentic, which means that their identity and their integrity have not been inadvertently or maliciously compromised, and that they are what they actually are what they purport to be (Council on Library and Information Resources, as cited in Duranti, 2005).

Further More:

> To an archivist, an authentic record does not have to be an accurate record; however, although it is true that an authentic record is as reliable and accurate as it was when first generated, this is not the same thing as saying that authenticity ensures that the content of a record at the point of its creation is accurate. Thus authenticity alone does not automatically imply that the content of a record is reliable or accurate (Roeder et al., 2008, p.14).

In an archival context an accurate record is one that contains correct, precise and exact data. When a record is created and used in a business process there is an assumption that inaccurate records harm business interests (Roeder et al., 2008) and accordingly an accuracy measure should be taken to ensure the content of a record is accurate.

Further the Trustworthy Repositories Audit and Certification: criteria check list, (TRAC) by OCLC and CRL (2007) states under requirement B1.4:

Repository's ingest process verifies each submitted object (i.e.,SIP) for completeness and correctness.

This basically means that the archival institution needs to process and check the deposited digital information when it is ingested to determine and qualify correctness and completeness. It further means that this process may include activities ranging from manual, human checking to the application of different tools for the automatic checking of data for any possible problems that degrade the correctness and completeness of the ingested data.

TRAC: It is a product of a collaborative effort between RLG and the National Archives and Records Administration to specifically address digital repository certification with a goal of developing criteria to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections (OCLC and CRL, 2007).

According to Digital Curators web site,

TRAC:

- Provides tools for the audit, assessment, and potential certification of digital repositories
- Establishes documentation requirements required for audit
- Delineates a process for certification
- Establishes appropriate methodologies for determining the soundness and sustainability of digital repositories (Digital Curation Center, 2010).

From the above discussions we can see that the digital objects an archival institution will preserve must be checked for accuracy, completeness and correctness.

Within different literatures we find that accuracy, correctness and completeness, are listed as the major data quality dimensions (Bovee et al., 2003), (Redman, 1996), (Batini and Scannapieco, 2006). In order to measure and improve the quality of data, different data quality activities can be performed. One of the most important digital related data quality activities are integration and the matching of two files in which inaccurate records are included, in order to find similar records that correspond to the same real-world entity through an approximation method approximate method (Batini and Scannapieco, 2006).

Batini and Scannapieco in their book (data quality, 2006) stated,

> Due to heterogeneous schemas, and to possible errors in data entry and update processes, objects happen to have different representations and values in distinct databases. As a consequence, a loss of a clear identity may affect objects, thus compromising the possibility of reconstructing information sparse in distinct sources. Object identification techniques aim at repairing this loss of identity using context information available on the similarity of objects' representations in terms of tuples, hierarchical relations, and XML files (Batini and Scannapieco, 2006, P.146).

Object identification is the data quality activity needed to identify whether data in the same source or in different ones represent the same object of the real world.

In most literatures two kinds of accuracy has been identified, syntactic accuracy and semantic accuracy.

Syntactic accuracy is the closeness of a value *v* to the elements of the corresponding definition domain D while semantic accuracy represent the closeness of the value *v* to the true value *v'* (Batini and Scannapieco, 2006). Accordingly if *v*=Karl and *v'*=Jensen, even if they are different, we can say *v* is syntactically correct since it is an admissible value in the domain of person's name, but in semantic accuracy we have to know if it is the right name, and it lends itself to the correctness measure.

While syntactic accuracy can be measured using functions, called comparison functions, that evaluate the distance between *v* and the values in D. Edit distance is a simple example of a comparison function, taking into account the cost of converting string *s* to a string *s'* through a number of character insertions, deletions, and replacements. For example we can look from table 1 the accuracy error of movie 1 on the Title value is a syntactic accuracy error. As the correct value of The Pursut of Happyness is The Pursuit

of Happyness,the edit distance between the two values is equal to 1.on the other hand if we swap the values of directors' names for movies 1 and 2 results in semantic accuracy error, as this results in wrong directors of the movies. Semantic accuracy is a complex one to identify and measure. In a more general context, a technique for checking semantic accuracy consists of looking for the same data in different data sources and finding the correct data by comparisons. This latter approach also requires the solution of the object identification problem, i.e., the problem of undestrstanding whether two tuples refer to the same real-world entity or not (Batini and Scannapieco, 2006), so object identification is an important step towards measuring and identifying semantic accuracy.

ID	Title	Director	Genre	Year
1	The Pursut of Happyness	Gabriele Muccino	Drama	2006
2	Titanic	James Cameron	Romance	1997
3	Dances with wolves	Kevin Costner	Adventure	1990

Table 1.1 Accuracy example

Furthermore object identification techniques are used to improve completeness, resolve inconsistencies, and eliminate redundancies during data integration process (Madnick, et.al, 2009).

We have introduced and discussed the importance of object identification as a major step towards ensuring that the digital holdings of an archival institution is both accurate and correct, and as such will focus this study to look at how to implement object identification on an archive that has a responsibility to take care of Norwegian public administration digital records in created and maintained in Norwegian archive format known as (Noark).

1.4 Statement of the problem

With the introduction of networks and the internet, XML standard becomes common. The current and the future of data transfer standard is XML, it is increasingly popular especially for data published on the web and data exchanged between organizations (Battini and Scannapieco, 2006). Despite the fact that, the increasing popularity of XML

standard, the implementation of object identification for such type of data is not common. According to Milano et al.(2005) and Weis and Naumann(2005),while there is extensive research in the realm of object identification in relational data, there is little attention given to other data models, such as semi-structured data, represented in XML standard. As a result of this fact, this paper will focus on an object identification of XML data.

How can an archive identify whether the data it is depositing from same or different sources represent the same real world entity?

1.5 Purpose of the study

The purpose of this thesis is to study the implementation of an object identification technique on representative archival data.

1.6 Aims and Objectives

- ✓ The research aims to investigate the Noark documentation and metadata catalogue to understand the structure of Noark 5 and structure of data,
- ✓ Identify the requirements of object identification
- ✓ Develop a prototype of an object identification technique,
- ✓ To test the chosen object identification technique on a sample data prepared on Noark-5 data format.
- ✓ Measure the effectiveness of the object identification approach

1.7 Research Questions

1) What are the functional requirements for automatic object identification techniques for Noark data?
2) How effective is the prototyped object identification technique?

1.8 Significance of the study

Organizations needs to make several decisions on their business day to day activities, and they made these decisions helping by data or information, in this regard Data quality is central to organizations meeting their goals and targets.

This thesis will primarily help in creating awareness of data quality, in particular for the archival community as it takes the Noark case.

It also helps those who want to do further research on the subject area, as it covers a good portion of the literature and mentions the different concepts in relation to data quality.

Chapter 2- Literature Review

2.1 Introduction

Following the transition of the world from an industrial economy to an information economy, many believed Information will be a key competitive advantage and companies were heavily involved in implementing information technologies. However advanced technology alone does not promise the success of a business, but the quality of information flowing through those technologies also matters most: High quality information is also necessary to enhance the competitiveness of a business. As Redman stated, "if information technologies are the engines of the information age, then data and information are the fuels" (Redman, 2001, xiii). It is this fuel that guides organizations to successful accomplishment of their business. High quality information benefits firms, whereas poor quality information hurts them (Huang et al., 1999, p.4). To strengthen the importance of information Eckerson (2002), mentions companies' main asset is information, and they compete on their ability to absorb and respond to information.

If we say so about the importance of information, actually the basic units of information is data, unless we make sure the quality of the basic unit "data" is ensured, we cannot talk about information quality; without highly accurate data, information quality cannot be achieved (Olsen, 2003). The Data Warehousing Institute report by Eckerson (2002) stated the importance of data as "if information is the currency of the new economy, then data is a critical raw material needed for success". Data is the fuel we use to make decisions. Everyone uses data; many large organizations are nothing but data processing engines, insurance companies, banks, financial service companies engage in processing data. Other organizations may appear to be less involved with information systems because their products or activities are not information specific, but looking in depth we found that most of their activities and decisions are driven or guided by information systems. Manufacturing organizations produce and ship products; however data drives the processes of material acquisition, manufacturing work flow, shipping, and billing.

The above introductory statements shows that data or information is an important asset to the success of one's business, and advocates the need for change of attitude in organizations, Lee et al., in their book Journey to data Quality claims the importance of data centric approach to solve the quality problems of organizations,"Most organizations have focused too narrowly on the systems side of the problem to the detriment of the data side of the problem. Addressing the data quality problems will

lead to increased capability to share information across the organization. Without this focus most systems solutions will be prone to failure" (Lee et al, 2006).The above statement is in strong agreement to the believe of this study, the quality of data is critical factor to accomplish goals and meet targets in organizations, for organizations to keep their data quality, they need different tools and methods, and this study looks part of the data quality aspect Object Identification.

2.2 Impact of poor information quality

The negative impact of poor data quality on companies has been documented in industry (Redman, 1996), (English, 1999). Data quality problems result in lower customer satisfaction and increased costs. In terms of economic aspects poor data quality generates maintenance and repair costs and beyond the economic aspects it can affect customer satisfaction, reputation and strategic decisions (Goasdoue, et. al.,n.d.). Olsen on his book Data Quality: The accuracy dimension, mentioned the following examples of impact of poor data quality:-

1) Transaction Rework costs:-many organizations have entire departments that handle customer complaints on mishandled orders and shipments. When the wrong items are shipped and then returned, a specific, measurable cost occurs. There are many data errors that can occur in this area: wrong part numbers, wrong amounts, and incorrect shipping addresses, to name a few. Poorly designed order entry procedures and screens are generally the cause of this problem.
2) Costs incurred in implementing new systems: one of the major problems in implementing data warehouses, consolidating databases, migrating to new systems, and integrating multiple systems is the presence of data errors and issues that block successful implementation. Issues with the quality of data can, and more than half the time does increase the time and cost to implement data reuse projects by staggering amounts.
3) Delays in delivering data to decision makers: many times you see organizations running reports at the end of time periods and then reworking the results based on their knowledge of wrong or suspicious values. When the data sources are plagued by quality problems, it generally requires manual messaging of information before it can be released for decision-making consumption. We can measure the wasted time of people doing this rework but the poor quality of decisions made cannot be measured.
4) Lost Customers through poor service: customers that are being lost because they consistently get orders shipped incorrectly get their invoices wrong, get their payments entered incorrectly or other aspects of poor service represent a large cost to the corporation.

5) Lost production through supply chain problems: whenever the supply chain system delivers the wrong parts or the wrong quantity of parts to the production line, there is either a stoppage of work or an oversupply that needs to be stored somewhere. In either case money is lost to the company (Olsen, 2003).

2.3 The challenge towards data quality

Despite losing a lot of assets due to poor data quality, it is seldom to see companies taking the necessary measure. Among the major reasons usually given are, not accepting the truth, that means companies do not accept they have the problem, overlooking or ignoring the impact that poor data quality can have in their businesses and not putting resources to take the necessary measures, executives do not want to expose their organization. Part of the problem is that "most organizations overestimate the quality of their data and underestimate the impact errors and inconsistencies can have on their bottom line" (Eckerson, 2002). Olsen (2003), strengthens this idea understanding executives too often affected by their natural behavior and do not want to believe they have problems in the first place. Moreover, senior executives may be further reluctant if the organization is performing adequately, particularly if calling attention to the problem would affect the corporate image (Lee, et. al., 2006).

The other challenge is that even if the management accepts their data has problem, they might not take it as a series problem. It is difficult to make the management believe the data quality problem. Senior management may not realize that poor data quality has a direct impact on the organization's performance. "Executives are often unaware of data quality problems, or they tend to believe that the information technology department can handle the problems without the allocation of critical time and resources from the top level" (Lee, et. al., 2006).

Executives are reluctant to commit scarce resources to an apparent no problem unless there is a defensible cost/benefit justification to show the costs of neglecting the data quality problem and the economic benefits of initiating the data quality project.

Looking the other dimension of the challenge, we find the technical difficulties of the data quality problem. Technologies are changing frequently, the dimension and types of data is increasing, the usual architecture of exchanging information is also changing, so do the data quality methods,techniques,tools needs to keep the pace with all those changes.

> In addition to solving existing problems, the community will face new challenges arising from ever-changing technical and organizational environments. For example, Most of the prior research has focused on the quality of structured data, in recent years; we have seen a growing amount of semi-structured and

unstructured data as well as the expansion of datasets to include image and voice. Research is needed to develop techniques for managing and improving the quality of data in these new forms. New ways of delivering information have also emerged. In addition to the traditional client-server architecture, a server-oriented architecture has been widely adopted as more information is now delivered over the internet to traditional terminals as well as to mobile devices" (Wang, et. al., 2009, p.17).

2.4 Benefits of good data quality

According to Redman (2001,p.17),the estimated cost of poor DQ is at least two percent of revenue, which does not include the invisible loss of corporations' reputation and customers' satisfaction, impacts can also include operational inconvenience, poor decision making, and in extreme cases, business closings. Experts have estimated the cost of poor information quality at from 15 to 25% of operating profits (Eckerson, 2002). It has been also estimated that the data quality problems can cost companies as much as 8-12 percent of revenue in industry (Redman, 1996). To worsen the situation is often the magnitude of these problems may be unclear,unquantified,or unrecognized by managers who have become accustomed to accepting the costs associated with data quality problems as the "normal cost of doing business"(English, 1999). We can see that the benefits of good data quality comes with by avoiding the above mentioned costs of poor data quality, if organizations do an effort on improving their data quality, they could minimize those figures, and also create a better environment for the overall performance of their organization. Better-quality information systems will reduce the cost of, and accelerate the completion of, steps in evolving the organization to newer business models, changing a corporation's business and operating systems to a base of high- quality data makes changes occur faster, at lower cost, and with better-quality outcomes (Olsen, 2003).

For non-profit organizations, like universities, libraries, archives, the importance of Data Quality cannot be overstated as well. High quality information not only helps them make sound decisions, but also adds value to the services they provide. The general nature of all data quality problems and the above examples is that data quality issues have caused people to spend time and energy dealing with the problems associated with them (Olsen, 2003).

2.5 Possible sources of poor data quality

What can go wrong? The U.S. Department of Defense (DoD) on CyKana, et.al.(1996) has a comprehensive set of guidelines on data quality management and Four types of causes for data quality problems are identified(1)process problems refer to problems in data entry,assignment,execution,and data exchange,(2)system problems come from

undocumented or unintended system modifications, or incomplete user training and manuals,(3)policy and procedure problems refer to conflict or inappropriate guidance, and(4)database design problems occur as result of incomplete data constraints.

The data warehouse institute report on data quality by Eckerson(2002),accept the fact that the sources of poor quality data are myriad and identified data-entry processes, as producing the most frequent data quality problems, and systems interfaces. Following are examples of the possible sources of data quality:

- Lack of validation routines. Though many data-entry errors can be prevented through the use of validation routines that check data as it is entered into web, client/server or terminal-host systems.
- Valid but not correct. But even validation routines cannot catch typos where the data represents a valid value.
- Mismatched syntax, formats and structures. Data-entry errors are compounded when organizations try to integrate data from multiple systems. These different systems may use different formatting, structures, so either a data cleansing or ETL tool needs to map these differences to a standard format before series data cleanup can begin.
- Spider web of interfaces.beacuse of the complexity of system architectures today, changes to source systems are easily and quickly replicated to many other systems, both internal and external. Most systems are connected through a spider web of interfaces to other systems. Thus; changes in source systems can wreak havoc on downstream systems if adequate change management processes are not in place.
- Lack of referential integrity checks. It is also true that target systems do not adequately check the integrity of the data they load.
- Poor system design. Source or target systems that are poorly designed can create data errors (Eckerson, 2002).

2.6 Evolution of Data Quality

The literature on evaluating and improving data quality is relatively new, dating back only to the mid-1970s with work done for the energy information administration (EIA) on the quality of a set of surveys of the nation's energy resources. The topic of data quality reemerged as an important issue in commercial and government organizations in the late 1980s. Commercial organizations were driven by strong competitive pressures to reengineer and improve their business processes, and data began to be seen as a key asset in this drive.

It was the wide implementation of computer networking, the introduction of relational databases and following these changes access to a large number of databases brought the question on the quality of data in various sectors, as Galway and Hanks explained,

> Following the explosive increase in computer networking, people had gain access to a much wider array of databases, leading to an increased awareness that much of the available data was of questionable if not poor quality. This has been highlighted by data-quality studies of scientific, medical, justice, and business databases and by the occurrence of some very expensive business mistakes (Galway and Hanks, 1996, p.5).

The justice sector started to give attention to the problems caused by poor-quality records in administrative processes, the 1979 scenario of court ruling regarding the bail proceeding reflected the problem of low-quality criminal records(Lee,2003).

Another scenario is in the military. A good example is the "Logistics Information Requirements and Quality" project sponsored by the U.S. army in 1996, which examined data quality problems in army logistics in hope of better supporting decision-making (Galway and Hanks, 1996). The report addresses problems in the quality of army logistics data and information, the causes of those problems, and potential fixes. It is widely perceived in the army that severe problems exist with the logistics data that provide the basis for many important army decisions.

In 1990s more systematic research regarding data quality was initiated in the academic world. Prominent contributions to the academic literature in the field have come from the work of Redman and his colleagues at AT&T, and from the program in Total Data Quality Management (TDQM) at MIT, directed by Wang, Redman (1992), English (2001) as cited in(Lee,2003).

2.7 Definitions of data quality

2.7.1 Data vs. information

Different literatures define data as raw materials for information, or set of facts. Since data in this study involves a computerized database, and such data are of greatest importance in quality control, the definition used by the database community is appropriate for this study. Data "known facts that can be recorded and that have implicit meaning" whereas A "database is a collection of related data" (Elmasri and Navathe, 2004). For example we may have stored or recorded the names, telephone numbers, and address of employees on a hard disk. This set of related data with implicit meaning forms a database. In defining a database, a data model is used to describe the structure of the database. The structure of database refers to "the data types, relationships and constraints that should hold on the data" (Elmasri and

Navathe, 2004). Redman (2001), adopted this view and defined data as consisting of two interrelated components, ”data models” and “data values.” Data models define what the data are all about, and different models reflect different aspects of the real world. A data model involves entities, attributes and relationships. An entity represents a real-world object or abstraction such as employee, customers or products. Attributes and relationships describe pertinent features of the entities (Redman, 2001).” data values” are assigned to attributes in the data model for specified entities (Redman, 2001, p. 71).

Galway and Hanks further explains the difference, “data or data elements, are specific entries in a database or an information system, information is the combining of different pieces of data to produce new quantities that provide insight into the processes producing the data” (Galway and Hanks, 1996, p.2).

Considering the field of data quality, which uses a wide spectrum of possible data representations, we can find different ways of classifying types of data, the classification used by batin and scannapieco related to the purpose of this study and presented below:

1. Structured, when each data element has an associated fixed structure. Relational tables are the most popular type of structured data.
2. Semi-structured, when data has a structure which has some degree of flexibility.XML is the markup language commonly used to represent semi-structured data. some common characteristics are (i) data can contain fields not known at design time, (ii) the same kind of data may be represented in multiple ways;(iii)among fields known at design time, many fields will not have values.
3. Unstructured, when data are expressed in natural language and no specific structure or domain types are defined (Batini and scannapeico, 2006).

R(FirstName,LastName,club,position)

Marlon	MZ	MU	Defender

Marlon	Mendez	Manchester United	Defence

a) Two tuples

<Player> <FirstName>Marlon</FirstName> <LastName>MZ</LastName> <Club>MU</Club> <Position>Defender</Position> </Player>	<Player> <FirstName>Marlon</FirstName> <LastName>Mendez</LastName> <Club>Manchester United</Club> <Position>Defence</Position> </Player>

b) Two XML records

Fig 2.1 Examples of Matching Objects in different data models

Among the above forms of data, the focus of this study is semi-structured data, which is expressed using the XML markup language, as the importance of XML, as a data exchange standard in organizations become prevalent.

Most literatures agree on the multifaceted nature of data quality, and they give different measuring dimensions to talk about data quality, and the frequently mentioning definition is “fitness for use”, being useful for the task at hand.

2.8 Data Quality Dimensions

Choosing dimensions to measure the level of quality of data is the starting point of any DQ-related activity. A vast number of bibliographic references address the definition of criteria for measuring data quality. Criteria are usually classified into quality dimensions (Olsen, 2003),(Redman, 1996),(Redman, 2001),(Lee, 2006),(Batini and Pannacaipaca,2006).None of the data quality dimensions is complete by itself, and many dimensions at times are overlapping, the following are some of the most common data quality dimensions found in the literature:

Accuracy: closeness between a value v and a value v' considered as the correct representation of the reality that v aims to portray. Olsen on his book Data Quality: The accuracy Dimension, claim accuracy dimension is the foundation dimension and argues that if the data does not represent the true facts, all other dimensions are less important (Olsen, 2003).

Completeness: it refers to the extent to which the expected attributes of data are provided. It can be viewed from at least three perspectives: schema completeness, column completeness, and population completeness. By schema completeness, we mean the degree to which entities and attributes are not missing from the schema. By column completeness, we mean the degree to which there exist missing values in a column of a table, and by population completeness, we mean the degree to which members of the population that should be present are not present.

Consistency: data are consistent if they respect a set of constraints. The data in the enterprise should be in synch with each other.

Relevancy: whether the data is useful for the task at hand.

Timeliness: reflects how up-to-date the data is with respect to the task for which it is used. A metric to measure timeliness has been suggested by Ballou et al., (as cited in Pipino et. al., 2002) accordingly timeliness be measured as the maximum of one of two terms: 0 and one minus the ratio of currency to volatility.volatitlity refers to the length of time data remains valid; delivery time refers to when data is delivered to the user; input time refers to when data is received by the system; and age refers to the age of the data when first received by the system.

$$\text{Max}\{0, 1 - \text{currency} / \text{volatility}\}$$

We have seen data quality is a multi-dimensional concept and it has associated dimensions. These dimensions are usually defined in a qualitative way and do not provide quantitative measures, so we need metrics associated with dimensions to measure quantitatively (Batini and Scannapieco, 2006). Data quality assessments are made both subjective measurements and objective measurements. Subjective data quality assessments based on the needs and experiences of stakeholders (Ballou,et. al.,and Wang, as cited in Pipino,et. al.,2002),whereas objective assessments further classified as task-independent metrics is used without the contextual knowledge of application ,and task dependent metric depend on the organizations business rules, company and government regulations, and constraints(Pipino et. al.,2002).

2.9 Functional Forms

Objective assessments to data quality can be achieved by developing metrics. Pipino et al. (2002), proposed a three functional form for the objective assessment of data quality in organizations.

Simple ratio: the simple ratio measures the ratio of desired outcomes to total outcomes.

Min or Max Operation: to handle dimensions that require the aggregation of multiple data quality indicators (variables), the minimum or maximum operation can be applied.

Weighted average: An alternative to the min operator in a multivariate case, if the company knows the importance of each variable to the overall evaluation of a dimension.

2.10 Data quality activities

We have talked about the importance of data quality and agreed that organizations need to improve their data quality. So what can be done to keep the quality of data? "A data quality activity is any process we perform directly on data to improve their quality" (Batini and Scannapieco, 2006). It requires a multi-strand approach that addresses a number of activities from manual activity to an automated or computerized activity. Accordingly, a manual data quality activity can be checking the right address when our sent email bounce back, while a matching of two files which included inaccurate records with the goal of finding similar records which belong to the same real world entity (Batini and Scannapieco, 2006). Data profiling, normalization, and data integration are some of computerized data quality activities as mentioned in the literatures Goasdoue et al., (n.d.), Batini and Scannapieco (2006), Loshin (2007).

Despite the range of data quality activities Batini and Scannapieco (2006), confirm us that Object identification and data integration are of crucial importance in current business scenarios and also covered broadly in research and industrial perspective.

"Data integration refers to the organization's inventory of data and information assets as well as the tools, strategies and philosophies by which fragmented data assets are aligned to support business goals"(Information management,2011),it is concerned with combining data from various sources into one consistent stream providing a single view. As this paper is targeted to an object identification data quality activity, it will be discussed thoroughly in the following section.

2.11 Object Identification

We call object identification the data quality activity needed to identify whether data in the same source or in different sources represent the same object of the real world.

Record linkage is used when the matching activity is performed on simple structured data, for example (Ravikumar & Cohen, 2004). Object identification is an evolutive term for record linkage, and deals with complex structured data and XML documents where objects of the real world are represented, in a wider spectrum of structures than simple structured data (Batini & Scannapieco, 2006).

When I create a bibliographic database from my reference lists in this paper, I need to determine which citations refer to the same paper in order to avoid duplication. Merging of multiple databases needs first to determine which records represent the same entity and should therefore be merged.

An entity or object is an abstraction of real world thing with characteristics that can be expressed with its attributes. For example a person object or entity may have attributes of name, address, date of birth, social security number. An entity might be a business, a person, or some other type of unit that is listed (Winkler, 2006).

Poor data quality in a single database produces poor service quality and economic losses. Poor data quality referring to the same types of objects(e.g.,persons,businesses and portion of territory) in different databases yields poor results in all applications(e.g.,queries,transactions and aggregations) that access the same objects in the different databases. This type of access is typical of many Government / Business / Citizen-to-Government / Business / Citizen interactions.

For example, “to discover tax frauds, different agencies can cross-check their databases in order to search for contradictions or correlations among data: this is possible only if data referring to the same object can be identified” (Batini & Scannapieco, 2006).

Agency	Identifier	Name	Type of Activity	Address	New York
Agency 1	CNCBTB765SDV	Meat production of John Ngombo	Retail of bovine and ovine meats	35 Niagara	New York
Agency 2	0111232223	John Ngombo canned meat production	Grocer's shop,beverages	9Rome Street	Albany
Agency 3	CND8TB76SSDV	Meat production in New York state of John Ngombo	Butcher	4,Garibaldi Square	Long Island

Table 2.1 Example of Object identification problem from (Batini and Scannapeica, 2006, p.97)

From the above table we can see that, agency 1 and agency 3 shares a common domain but due to some typo error they seem different.

In business scenario for example:

> Integrating all the reviews of restaurants from the Zagats's restaurants webpage with the current restaurant health ratings from the department of Health's website, to integrate these sources requires comparing the objects from both sources and identifying which restaurants are the same (Tejada et al., 2001, p.607).

2.12 Historical perspective

The term has different names in different literatures, entity resolution (Wang & Madnick, 1989), (Talburt et al., 2005), also known as record linkage (Winkler, 2006) and object identification in (Tejada et al., 2001), (Batini and Scannapieca, 2006). The term record linkage is mentioned for the first time in (Dunn, 1946). It was a time where computer applications have been developed to automate different activities in different sectors, like administrative activities, demographic studies, health experiments, and epidemiological analyses, and data often result from merging of different sources, created and updated at different times, by different organizations or persons and the fact that merging data produces new data of potentially higher value, since properties that are merged can be related with new types of aggregations,analysis,and correlations. In 50's and 60's, data was represented in files, records, and fields, and terminology that justify the original term record linkage as the activity that results in the integration of information from two or more independent sources. In a number of cases, the files are too big to consider every pair and Newcombe (1962) came up with a

method to reduce the number of pairs to be considered by only considering pairs that agreed on a characteristic (Winkler, 2006), i.e., frequencies of occurrences of values in strings and decision rules for matching and non-matching records, (Newcombe, 1959, as cited in Batini and Panncapieca, 2006). Later a more formal method was introduced by Fellegi and Sunter (1969) provided a formal mathematical model for ideas that had been introduced by Newcombe, 1959, 1962(Winkler, 2006). The idea is to classify in a product space AXB from two files A and B into M, the set of true matches, and U, the set of true non-matches. They used ratios of probabilities. Following this a number of experiments were done considering large amount of data, with various degrees of trustworthiness and accuracy and from different application areas (Batini & Scannapieca, 2006), (Winkler, 2006). In recent years with the introduction of different technologies, increased type of data, and a different way of exchanging information are forcing new strategies, techniques and metholdologies,"New techniques have been proposed that extend the linkage activity from files to more complex structures. Such techniques also try to exploit knowledge on the application domain to produce more effective decision procedures" (Batini and Scannapieca, 2006).

2.13 Techniques for object identification

Well we have a data quality activity, Object Identification, which can help identifying objects which belong to the same real world entity. But how does it work exactly, what does it uses? It is an object identification technique which does the work, which helps the object identification to identify the same objects. "Techniques correspond to algorithm, heuristics, knowledge-based procedures, and learning processes that provide a solution to a specific DQ problem, or as we say, to a data quality activity" (Batini & Scannapieca, 2006).

Object identification Techniques used either to improve efficiency or effectiveness. According to Batini and Scannapieca (2006), Weis et al. (2006), two sets of techniques have been identified, techniques that mainly focus on efficiency and techniques that focus on effectiveness. Those techniques focusing on efficiency usually tries to reduce the number of pair wise compassion so that saving computation time, while on the other hand the effectiveness techniques focus on improving quality of duplicates (Weis et al., 2006).

The data quality problem has been investigated by different disciplines, Ravikumar and Cohen (2004), tells us that the different communities have been looking the problem in the context of their study and formulates different techniques. For example, Knowledge intensive approaches has been used in the database community, later further development of string edit distance as general-purpose record matching scheme proposed by Monge and Elkan,in the AI community the application of supervised learning to the record-linkage task-for learning the parameters of string-edit distance

metrics, and combining the results of different distance functions. More recently, probabilistic object identification methods have been adapted to matching tasks. In statistics, a long line of research has been conducted in probabilistic record linkage, largely based on the seminal paper by Fellegi and Sunter (Ravikumar & Cohen, 2004).

Techniques developed for dealing with the object identification problem strictly depends on the type of data used to represent objects. It depends how good the data structure to represent semantics is. While Simple structured data correspond to traditional files and was poor in representing semantics, but later with the development of Relational database management systems made it possible to assign semantics to such data in terms of domains, keys, functional dependencies and constraints. With the introduction of networks and the internet XML standard becomes common and different object identification techniques are being implemented.

It is on Batini and Scannapieca (2006), work that object identification techniques have identified and discussed thoroughly. Most literatures have identified three major categories of techniques for object identification on the basis of the underlying research paradigms:

Name	**Technical Area**	**Type of data**
Fellegi and Sunter and extensions	Probabilistic	Two files
Cost-based	Probabilistic	Two files
Sorted Neighborhood and variants	empirical	Two files
Delphi	empirical	Two relational hierarchies
DogmatiX	empirical	Two XML documents
Intelliclean	Knowledge-based	Two files
Atlas	Knowledge-based	Two files

Table 2.2 Object Identification techniques from (Batini and Scannapieca, 2006, p.107)

Probabilistic techniques, based on the extremely relevant set of methods developed in the last two centuries in statistics and probability theory, ranging from Bayesian networks to data mining tools. Fellegi and Sunter is the first and by far the more established technique, and it is representative of probabilistic techniques.

Empirical techniques that make use in the different phases of the process of algorithmic techniques such as sorting, tree traversal, neighbor comparison, and pruning (Batini & Scannapieca, 2006). The sorted neighborhood method and its variants are representative of this category. The sorted neighborhood method works first it sorts all the entities based on a pre-selected key attributes an heuristics, and a fixed size window is moved from beginning of the list to the end, during which the first entry is compared to the rest of the list and matched based on distances(Yan et al.,2007).

Knowledge-based techniques, in which domain knowledge is extracted from the files involved, and reasoning strategies are applied to make the process more effective. DogmatiX is among the first techniques dealing with object identification in XML documents, and Delphi is among the first ones dealing with complex structured data.

Since the focus of this paper is an Object identification of xml documents, a through discussion will follow on such strategies.

While there is extensive research in the realm of Object Identification in relational data, there's little attention given to other data models, such as semi-structured data, represented in XML standard (Milano et al.,2005),(Weis & Naumann,2004).

What makes the object identification on XML documents different from Object Identification on relational data? The hierarchical and semi-structured nature of XML compared to the flat, well structured relational data complicates the object identification task on XML data (Batini & Scannapieca, 2006).

Accordingly, two difficulties exist in xml data object identification: 1) Object definition, i.e., which data values actually refer the object and which data values refer the description of an object among the nested xml elements.2) the structural diversity addresses the fact that, XML elements describing the same kind of object are not necessarily equally structured, which can be accounted for different representations of same objects, or differences allowed by schema,e.g.,multiplicities of elements.

This is also reflected on Milano et al. (2006), work; emphasizing that method of object identification on XML data should be able to address problems raised due to the structural flexibility in addition to textual errors, i.e., typographical errors:

> "XML documents often represent complex, nested data, and schema languages for xml allow great flexibility in how such values is represented inside a document.xml data representations may allow for optional values, and lists of values whose length is not known schema-wise. Functions for approximate XML

data comparisons must thus be able to cope both with errors at the level of textual data values and with structural flexibility" (Milano et al., 2006, p.1).

In the following section we will look some of the works done on XML data Object Identification, if we look most of the works; it is based on extracting the data and putting on a flat structure before proceeding to the comparison and classification.

Most proposals for XML object identification are structure oblivious, in the sense that they rely on some kind of flattening of document structure to perform comparisons (Milano et al., 2006).

In Sahinalp (2003), xml objects are flattened and compared using string comparison functions (as cited in Milano et al., 2006). In the dogmatix framework data is extracted from an xml document and stored in relations called object descriptions. Tuples of two object descriptions containing data with the same XPath are classified as similar or contradictory using string edit distance, and object descriptions similarity is assessed taking into account the number of similar and contradictory tuples (Weis &Naumann, 2005, as cited in Milano et al., 2006). Moreover it makes use of heuristics to explicitly specify which XML elements are objects, which XML elements describe the objects and a thresholded similarity measure among objects for classification (Weis & Naumann, 2005) and The approach in (Jagadish et al., 2002) is similar, but comparisons of two objects take into account also approximate similarity results of descendent objects as cited in (Milano et al., 2006)."The tree edit distance between trees representing XML instances is used as a metric to estimate the similarity, it considers some aspects of XML structures, like repeated and optional elements, and also restricts the sequence of allowed operations, reducing the cost to obtain the edit distance score"(Jagadesh et al.,2002,as cited in Goncalves & Mello,2007).

Moreover, as it is outlined above some of the techniques focus on effectiveness while others emphasize the efficiency of object identification.

Singla and Domingos (2007), shows that traditionally the object identification problem is solved separately for each pair of duplicate candidates and latter a transitive closure is applied to cluster the match and non-match of records in a file, i.e., an object

identification technique which follows such procedure needs to consider every possible pair of candidate for a match and that takes O(n2) number of matches, which is a big number even for a medium sized database.

The object identification problem has a search space dimension equal to the cardinality of AXB, given two sets of records A and B to be compared. And to reduce these search space methods like, blocking, sorted neighborhood and pruning can be used (Batini & Scannapieca, 2006). In Single and Domingos (2007), they used the technique of first clustering the dataset into possibly overlapping canopies using an inexpensive distance metric, and then applying an inference and learning algorithms only to record pairs which fall in the same canopy; this reduced the number of potential matches to at most 1% of all possible matches.

2.14 Measuring the effectiveness of Object Identification technique

The Object identification technique classifies a set of records into a matching M and non-matching U. This gives rise to the following results: False positives FPs for records declared as M while actually being U,and False Negatives FNs for records declared as U while actually being M.True Positives TPs,correctly identified as M and True Negativists TNs are correctly identified U. Mostly the effectiveness of object identification techniques are evaluated using the two widely known metrics Precision and Recall(Batini and Scannapieca,2006),(Weis et al.,2006). Recall measures how many true positives are identified in relation to the total number of actual matches.

$$\text{recall} = TP / M = TP / TP+FN$$

Precision measures how many true matches are identified in relation to the total number of declared matches, including erroneous ones (i.e.,FPs).

$$\text{precision} = TP / TP+FP$$

M	Actual match w.r.t.real world
U	Actual non match w.r.t.real world
FP	Declared match while actual non match
FN	Declared non-match while actual match

TP	Declared match while actual match
TN	Declared non match while actual non match

Table 2.3 Notation on matching decision cases from (Batini and Scannapieca, 2006, p.126)

The aim of object identification is to have a high recall and a high precision; despite the two measures are often conflicting goals.

2.15 The General Steps to Object Identification

The general steps to achieve an object identification technique follow the following procedures:

Prepropressing: which has a goal of working on data in order to standardize it and correct evident errors. Some of the activities include conversion of upper / lower cases to make data homogenous for comparison and replacement of null strings.

Search Space Reduction: given the search space AxB of the two files, find a new search space CcAxB to apply further steps. Three different methods can be applied to the search space reduction, namely, blocking, sorted neighborhood and pruning (or filtering).

Choose comparison function: choose the function(S) / set of rules that expresses the distance between records in C.approaches to solve the object identification problem make use of some kind of distance function to detect the similarity of two objects.

Apply decision model: choose the method for assigning pairs in C to M, the set of matching records the set of unmatching records, and P the set of possible matches

Verification: check the effectiveness of method if not satisfactory, go back to step 2

(Batini & Scannapieca, 2006).

The object identification techniques usually pass through experiments to learn the effectiveness and efficiency of the different algorithms, and in order to do so we need data to test with, these data can be found in two different ways; the first could be from Existing duplicates, a real world data that might contain duplicates, this option is not easy as it is difficult to identify where the duplicates lie, find and mark these duplicates. Another option is inserting duplicates; it is an artificial duplicate by contaminating a set of data assuming it was clean and non-duplicate. This also comes with a challenge how

to create them, how many and where to put them (Weis & Naumann, 2004). Such data contamination can be done through the use of tools prepared for such purpose, one of such tools is the xml Dirty Data Generator, which takes parameters for determining the number and type of duplicates.

Chapter 3-Methodology

3.1 Purpose of the research

The purpose of the research is to study the implementation of Object Identification on an archive. In order to do so the researcher will describe the structure of Noark 5 and its data and then the requirements of object identification and develop a prototype of the object identification technique. Finally an evaluation of the object identification prototype will be done in order to validate the object identification technique.

3.2 Research Approach

As setting up knowledge claims helps the researcher to have a certain assumptions about the learning mechanism (how he will learn and what he will learn) in the process of their inquiry, it is important. This research uses an Interpretivist philosophical perspective.Orlikowski and Baroudi (1991), following chua (1986), as cited in (Myers, 1997), suggests three categories of philosophical assumptions: Positivist, interpreteivist and critical. The believe that an objective truth exist is the positivist, while the interpretivist believe on the subjective understanding of phenomenon's, meanwhile the criticalist do the study on existing social systems and reveal any contradictions. "Interpretive research starts out with the assumption that access to reality (given or socially constructed) is only through social constructions such as language, consciousness and shared meanings" (Orlikowski, 1991). Attempts to understand phenomena through the meanings that people assign to them. As such in this study the researcher will be able to grasp the meaning of the phenomenon under study object identification of digital records and construct the knowledge to set up the functional requirements of object identification through a shared meaning of the field and apply in the context of the study which is an archive. Moreover as cited in Klein and Myers (1999), the increased importance of interpretive research on IS (Information System), the interpretive research can help to have a deep insight into information systems phenomenon including the management of information systems and information systems development, the focus of this study archive is one type of information system and development of an object identification system. To construct interpretive proposes that there are multiple realities, not single realities of phenomena, and that these realities can differ across time and place. What is taken to be valid or true is negotiated and there can be multiple, valid claims to knowledge. As such this research will find out one possibility of doing the object identification and on the context of an archive.

3.3 Research Design

In most literatures today, it is common to see three major research approaches, quantitative, qualitative and mixed methods research approaches. The quantitative approach usually believes on objective reality of social facts, there's objective truth while the qualitative approach sees socially constructed truth, while the mixed methods approach needs working on both qualitative and quantitative data. This study will use a qualitative interpretive research design approach. Interpretivism lends itself mainly to qualitative studies (Viliers, 2005). Mertens (1998) describes qualitative research as a naturalistic interpretive science which is multi-method in focus. Interpretive approach rely heavily on naturalistic methods which are qualitative (like interviewing and observation and analysis of existing texts) which ensure an adequate dialog between the researchers and those with whom they interact in order to collaboratively construct meaningful reality(Cohen & Crabtree,2006).

According to Mack et al. (2005), it allows greater spontaneity and adaptation of the interaction between the researcher and the study participant. Moreover the nature of qualitative research being flexible and allows an iterative approach to go back and forth matches with the importance of an iterative prototyping design to check back and forth the research question and design and implementation of the object identification approach as the researcher learns or understands the knowledge on the process of research.

The qualitative method of document analysis will be used and it helps to get an in-depth understanding of the structure of Noark-5, the type of data it holds, data structure, in order to do so the Noark-5 documentation manual and Noark-5 metadata catalogue will be analyzed and as necessary the web site of Noark will be consulted. After understanding the structure of Noark the object identification functional requirements will be identified and a prototype of an object identification system will be developed. Finally A qualitative evaluation of the object identification approach will be done on a Noark data in order to validate the prototype.

3.4 Research Strategy

This research will make use of a case study strategy. The case study aims at giving a "holistic account of the case and in-depth knowledge of the specific through rich descriptions situated in context. This may lead to an understanding of a particular phenomenon" (Pickard, 2007 p.86). Qualitative research methods suitable for data quality research include action research, case study, and ethnography (Myers, 1997). The objective of case studies is the conduct of research and the researchers themselves specify the research questions. These are efforts where the research questions are specified prior to the study by researchers who are observers or investigators rather

than participants. The study has given a clear objective by stating the research problem in the form of research questions.

There are different reasons why case study research is a useful approach to a research and when it is appropriate to use. first, the researcher can study the phenomenon of interest in a natural setting, learn about the state of the art, and generate theories from practice, second, the case method is a chosen strategy to answer "how" and "Why" questions (Yin, 2009),third, it is an appropriate way to research an area where research and theory are at their early, formative stages,i.e., in which few previous studies have been carried out and the problems are practical(Bonoma,1983,Roethlisberger,1977, as cited in Benbasat,1987). So case study is chosen as a research method for this study since the research conducted on object identification problem in an xml data is little, and more over there is not any done on an archiving context. Though the object identification problem has been indicated very practical.

Yin (2009,p.18) defined the case study research strategy as "an empirical inquiry that investigates a contemporary phenomenon within its real-life context", Benbasat(1987), put the importance of determining the unit of analysis prior to searching the sites,i.e.,the focus of the study,individuals,groups,or an entire organization,alternatively,it may be a specific project or decision. The unit of analysis for this study is "Object Identification", which is a contemporary phenomenon especially in a semi-structured data represented using XML standard; moreover the archive is a real-life context as the phenomenon takes place in databases, data warehouses. To strengthen the idea, According to Noor (2008), case study is preferred when the questions are targeted to a limited number of events or conditions and their inter-relationships, and this idea is also reflected in yin (1989) suggests that the term refers to an event, an entity, an individual or even a unit of analysis.

This research has chosen the Norwegian Archive (Noark 5) as a case; the reasons are its logical as the researcher is conducting his study in Norway and the strong archive and records management practice in Norway. An instrumental case study will be employed, as the interest of the research is the phenomenon of object identification applied to Noark 5 records and not any of the case study sites (Pickard, 2007).

3.5 Research Methodology

Prototyping is an application design and development methodology. Among the benefits of prototyping is saving costs as it contains basic feature of a system it helps to show things fast, it is an iterative approach by its nature which gives a flexible way of achieving goals by going back and forth on the requirements specification, design and implementation(Larson,1993).

Definition:-according to Webster's Dictionary the term prototype has three possible meanings:

1) It is an original or model on which something is patterned: an archetype.
2) A thing that exhibits the essential features of a later type.
3) A standard or typical example.

When to use prototyping: prototyping should be considered in situations where there is a need for experimentation and learning before commitment of resources to development of a full-scale system. Prototyping allows experimenting with different configurations and selecting the most effective way of doing things (Alavi, 1984).

"An entry condition required to start prototyping is the completion of a written requirements specification" (Ashvins Group), this written requirements specification needs to be validated through a requirements validation techniques, it helps to assure that the software requirements specification meets user's requirements. "Requirements validation techniques concerned with demonstrating that the requirements define that the customer really wants" (Somerville, 2006). According to Somerville prototyping is among requirements validation techniques:

Requirements reviews: systematic manual analysis of the requirements

Prototyping: using an executable model of the system to check requirements

Test-case generation: developing tests for requirements to check testability

Automated consistency analysis: checking the consistency of a structured requirements description (Sommerville, 2006).

The prototyping process consists of several iterative cycles; Floyd (1984) (as cited in Carr & Verner, 1997) describes the prototyping process as consisting of functional selection, construction, evaluation and further use. Those functions that are to be prototyped are selected and a prototype is constructed. This prototype is evaluated and the prototype is further used for outlining specification or as a part of the new system. While Nauman and Jenkins (1982), (as cited in Carr & Verner, 1997) characterize prototyping as a four step iterative procedure involving users and developers:

User's basic needs are identified; a working prototype is developed; the working prototype is then implemented and used; the prototyping system is revised and enhanced.

If electronic prototypes are desired, several tools are available to quickly create a representation of the software functionality. Prototyping techniques include the use of very high-level languages, database programming and prototype construction from reusable components.

Language	**Type**	**Application domain**
Smalltalk	Object-oriented	Interactive systems
Java	Object-oriented	Interactive Systems
Prolog	Logic	Symbolic Processing
Lisp	List-based	Symbolic Processing

Table 3.1 Prototyping Languages from (Sommerville, 2006).

In this implementation of a prototype of the object identification system, a high level programming language called Java will be used.Tha java language is chosen suitable for this implementation because it is an object oriented programming language, it is widely used and experimented language, the object identification system needs interaction between a human and the system, which the java language fulfills as it is desirable for applications in the domain of Interactive systems. It also works well with parsers such as DOM, SAX.

Prototyping applied in the software process has various types, according to Sommerville (2006):

A) Evolutionary prototyping: an approach to system development where an initial prototype is produced and refined through a number of stages to the final system
   The objective of evolutionary prototyping is to deliver a working system to end-users.the development starts with those requirements which are best understood.
B) Throw –away prototyping
   A prototype which is usually a practical implementation of the system is produced to help discover requirements problems and then discarded. The system is then developed using some other development process. The objective of throw-away prototyping is to validate or derive the system requirements. The prototyping process starts with those requirements which are poorly understood.

This study will implement an evolutionary prototype of the object identification approach to the Noark-5 records. The prototype will be an executable system that shows a demonstration on the processes of how object identification from the Noark-5 records can be done. It will take an xml document which contains a number of objects from Noark-5, identify the different objects and provide a match and non-match of objects. The evaluation of the object identification system will be done measuring the effectiveness level and accuracy level which can be computed in order to validate the object identification technique.

Chapter 4-Requirements Modeling

In this chapter the structure of Noark data is explained, the prototyping process is covered up including the prototyping plan, the functional requirement accompanied by the functional system components diagram, the system architecture diagram and flow chart diagram of the object identification system. Finally an algorithm of the object identification system presented.

4.1 Noark

Noark is Norwegian archive standard (Norsk arkivstandard). It was developed as a specification of requirements for electronic recordkeeping systems used in public administration. it has gone through different development stages, the first release Noark1 came in 1984,and later in 1987 Noark 2 and 1994 Noark 3 followed in order to cop up with modernization in line with technological advancements and expansions to the systems' information content and functionality.

It was the release of Noark 4 1999 which took the standard a major step further by specifying a complete electronic records management system, integrated with e-mail and general case handling systems. the strong pressure from both public sector and suppliers to bring about good solutions for integration between Noark-based systems and task systems coupled with the development of different national and international standards of relevance to electronic recordkeeping and archiving which are not covered adequately in Noark 4 brought the need for a project launch in 2005 to formulate Noark 5. The major specifications of Noark 5 cover the following:

- Record structure (the inner arrangement of the archive and hierarchical with several levels from top to bottom)
- Metadata (information which describes the documents in the archive, both physical and electronic documents)
- Functionality (the functionalities the system support)

4.1.1 Noark-5 inner core

The inner core will handle the organization's archive, i.e. the fonds documents that are received or produced as a result of the activities carried on by the organization.

Fonds documents come into the archive, i.e. they are archived, through document capture. The documents must be organized in a record structure which shows the link between the documents. This means the documents must be placed in the correct place in the archive. When documents are archived, they must be frozen for all further editing.

Documents capture also involves the documents being assigned metadata, i.e. information on the documents' content, context and structure. An important function of metadata is to maintain trust in the documents' authenticity over time. There must be no doubt that a document is genuine and that it was created by the person who claims to have created it. it must be possible for the archive structure to be administrated by those who have the necessary rights. It must for example be possible to move documents that have been incorrectly archived.

It must be possible for documents that have been archived to be retrieved quickly and securely, and it must be possible for both the documents and their context to be presented to users in a clear manner.

The inner core must also contain functions for preservation and disposal.organisations are neither required nor permitted to retain all their fonds documents for the same length of time, and it must therefore be possible to add rules for when such documents must be removed from the archive.

4.1.2 The archive structure

The models in Noark 5 are conceptual models which are intended to show the link between different metadata and between metadata and physical or electronic documents. The conceptual models in Noark 5 state something about how the information should be organized in principle. They also form the basis for the definition of data structures in connection with electronic communication, integration with other systems, and migration from one system to another and for transfer. The archive

structure outlined through the conceptual model in this section represents the main structure in Noark 5 and is obligatory for case records.

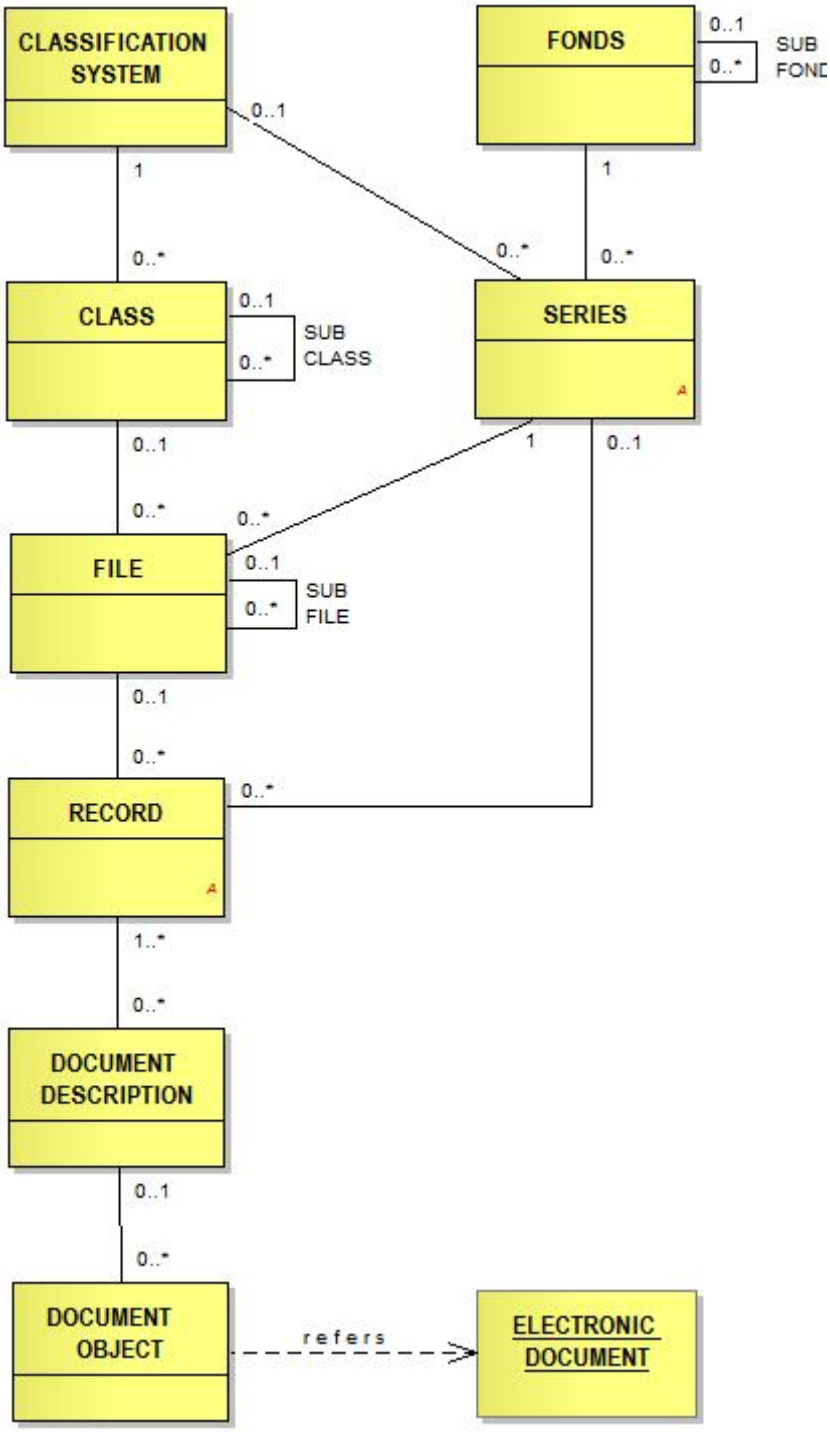


Fig 4.1 Conceptual model for Noark-5 from (Noark-5 standard for records management, 2009, p.39)

Simplified structure

In some task systems, there may be a need for a simplified structure in relation to case records. If there is no need to group records into files in a task system, the file level can be omitted. Similarly, the document description level can be omitted if a record always consists only of a single document and if this document will not occur in other records.

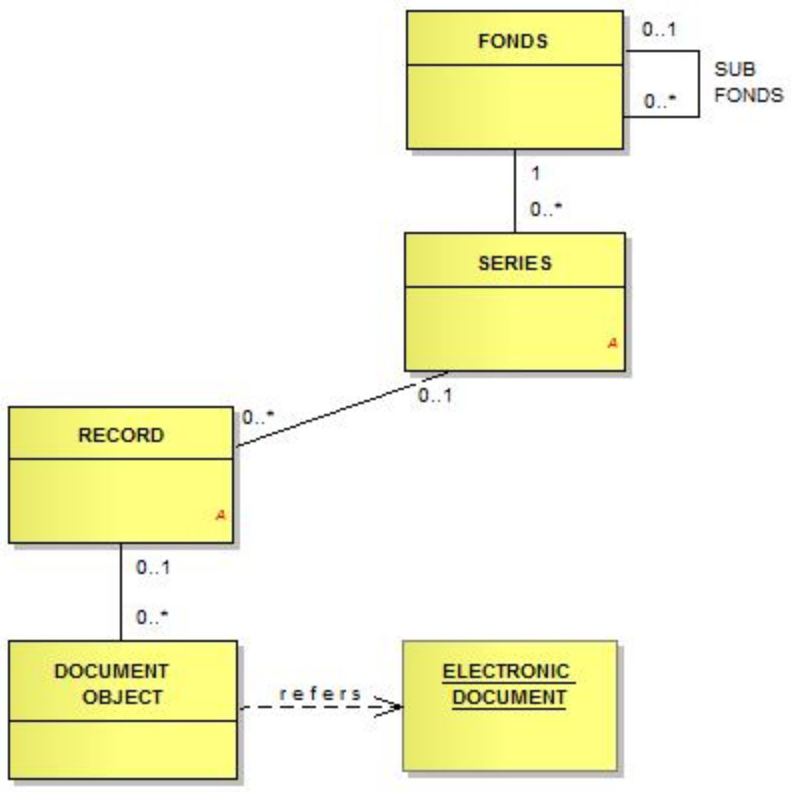


Fig 4.2 Simplified version of the conceptual model from(Noark-5 standard for records management,2009,p.41).

Noark- 5

Archive Structure for xml

1. nivå: arkiv(*)
2. nivå: arkivedel
3. nivå: klassifikasjonssystem(*)
4. nivå: Klasse
5. nivå: mappe(*)
6. nivå: registrering
7. nivå: dokumentbeskrivelse
8. nivå: dokumentobjekt

```
<arkiv>
    <arkivdel>
        <klassifikasjonessystem>
            <klasse>
                <mappe>
                    <registering>
                        <dokumentbeskrivelse>
                            <dokumentobjekt>
                            </dokumentobjekt>
                        <dokumentbeskrivelse>
                    </registering>
                </mappe>
            </klasse>
        </klassifikasjonessystem>
    </arkivdel>
</arkiv>
```

1. nivå: arkiv (*)
2. nivå: arkivdel
3. nivå: mappe (*)
4. nivå: registrering
5. nivå: dokumentbeskrivelse
6. nivå: dokumentobjekt

Fig 4.3 Noark-5 Archive structure for XML

```
<arkiv>
    <arkivdel>
        <mappe>
            <registrering>
                <dokumentbeskrivelse>
                    <dokumentobjekt>
                    </dokumentobjekt>
                </dokumentbeskrivelse>
            </registrering>
        </mappe>
    </arkivdel>
</arkiv>
```

Fig 4.4 Noark-5 alternative archive structure for XML

```
<records>
    <records section>
        <file>
            <registration>
                <document description>
                    <document object>
                    </document object>
                </document description>
            </registration>
        </file>
    </records section>
</records>
```

Fig 4.5 Noark-5 Archive structure for XML English translation

4.2 Prototyping

In this section in order to help structure the object identification solution, first we have put the prototyping process as a framework and based on this framework the prototyping plan is prepared, the functional requirements are identified and an executable prototype is developed and on the next chapter the evaluation of the implemented prototype presented.

4.2.1 Prototyping Process

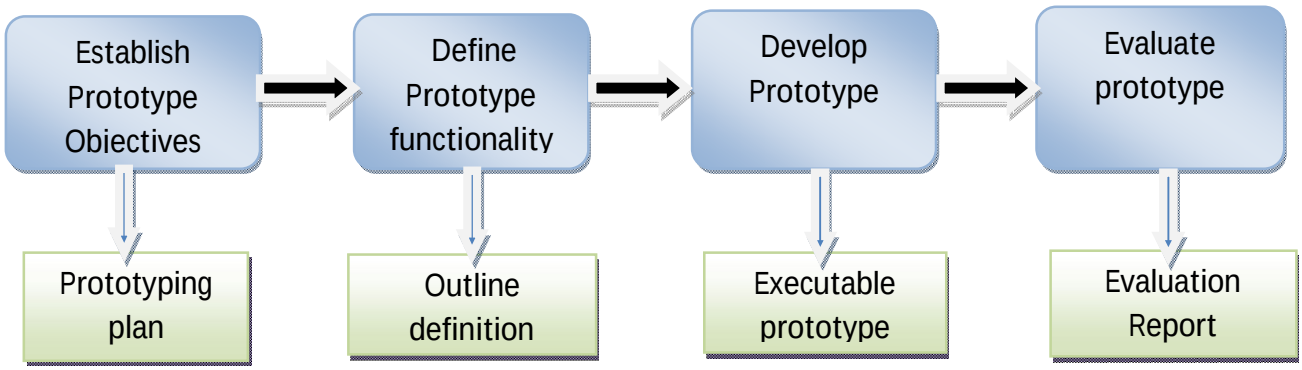


Fig.4.6 Prototyping process from (Sommerville, 2006)

4.2.2 Prototyping plan

What are we Prototyping?	• An Object Identification System *That checks if two objects are a representation of the same real world Entity • It reads an XML document which has a number of Objects within • It parses the objects • Select a description of the objects which can be used to help compare different objects • Compare a pair of objects using their description selection and a thresholded similarity measure • Classify objects into a class of Match and Non-Match

What do we want to know and why?	• The functional requirements of the Object Identification System • The effectiveness of the prototyped object identification system
How will we do it?	• The functional requirements of the system will be identified • The functional system component will be depicted in a diagram • System Architecture Diagram • Flow Chart Diagram • Algorithm implementation • Evaluation of the prototype Setting an experiment on an archive data which is formed as an XML document and artificially duplicated.i.e. A labeled data will be used which is a data with a priori knowledge of the matching status. • Evaluation report in terms of findings using the well known effectiveness evaluation metrics, recall and precision.

What method or approach will we use?	• A prototyping Approach • An Evolutionary prototyping approach, starting with the best known Functional requirements and eventually develop to a full system • An executable Prototype will be developed • An Object-oriented programming language Java will be used to develop the executable prototype of the Object Identification System

Table.4.1 Object Identification Prototyping Plan

4.3 Functional requirements

The functional requirement document also called functional requirement specifications, defines the capabilities and functions that a system must be able to perform successfully (Ofni Systems, 1999). In this case the capabilities and functions the object identification system performs.

Functional requirements should include: descriptions of data to be entered into the system, where the data is coming from / who is entering the data, descriptions of operations performed by each screen, descriptions of work-flows performed by the system, descriptions of system reports or other outputs (Ofni systems, 1999).

Accordingly the functional requirements of the Object identification system has been identified as a functional component / module, with the respective data entering to each

component of the system, where the data is coming from, description of the operations performed by each component module and their respective outputs.

Extract Component / Module

Extraction of XML document, in this task the system will be provided with the XML document which contains a number of objects suspect able for duplication within. This task can be accomplished in two ways: the first way is the system automatically extracts the xml document without the interaction of human. This is by giving the URL of the xml document within the Noark. Second, it can be done with the help of a staff personnel / Archivist by typing the location and name of the xml document for the system during run time.

We have to give the system a well-formed xml document so that it can forward it to the next process, otherwise the system cannot precede working on it.

In this step we have to make sure we are giving the system an xml document which has a possible duplicates within and need to be compared,i.e.,it makes no sense to compare objects of different real-world type, as they cannot be compared. A human is needed to involve by selecting a Comparable XML document.

For example the xml document we have prepared on section …is a possible candidate for the Object Identification System.

Duplicate candidate objects: -this is a step where the system identifies different individual objects find within the XML document which it has received from the Extraction component. A parsing using the DOM (Document Object Model) Parser will be done and individual / independent objects will be identified.

For our xml file created above the duplicate candidate objects identified by the system will be:

Object 1: First name:Jens,Last name:kohn,Phone: 34567898,Address: karlesgate.34,Department:Finance,Position: administrative head

Object 2: First name:Lans,Last name: krindahe,Phone: 76874534,Address: bronares.67

Department:IT,Position:Developer.

Conversion Component / Module: On this module the individual objects will pass through different stages to make them easy for the later steps of object identification, i.e., comparison and classification. The operations mainly done here are Object description selection and Object description generation.

Object Description selection: Every object has got characteristics which can help in identifying or characterizing that particular object. This characteristic is usually expressed by its attributes in a database context.i.e.the attributes of an object help to uniquely identify or characterize a particular object. At this stage the system will decide which attributes to choose as a representative identity of an object.

In this particular scenario, as it is not possible to accommodate every attribute as a description of objects, due to an overhead on comparison, it is been tried to narrow down the selection to a few powerful descriptions. Accordingly the description of objects consists of attributes which are listed as mandatory and with a String data type.

Object Description: The actual comparison of objects is done relying on the description of objects selected in the above step. In order to make things easier for comparison each object is prepared in such a way that their candidate descriptions will be put in an object description. The object description comprises of description of the objects as a relation of object description value and the name of that object description, relation (value, name).

For the above xml document we have created the system will make an object description of each objects as follows:

Object 1 {(Jens, First Name),(Kohn, Last Name),(Karlesgaet.34,Address)}

Object 2{(Lans, First Name),(Krindanhe,Last Name),(bronares.67,Address)}

4.3.1 Functional System Components

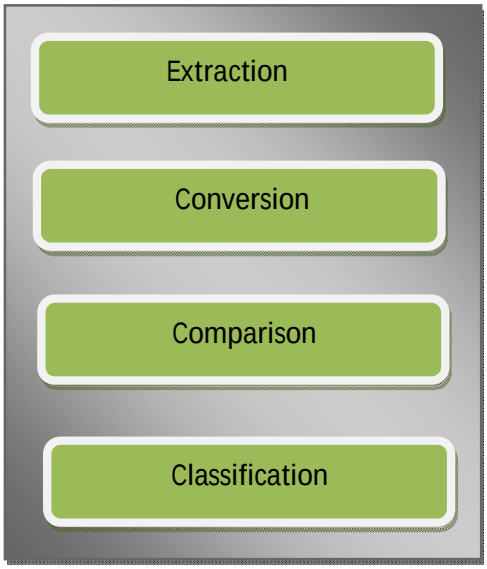


Fig 4.7 Object Identification functional System Components

Extraction: Get an xml document which contains candidates for object identification.

Conversion: prepare the xml document for further steps. Like identifying different objects, identify characteristics of the objects.

Comparison: do the actual identification of objects, i.e. comparing different objects based on their characteristic descriptions for similarity

Classification: categorize objects into match and non-match based on the similarity results.

Comparison Component / Module:-this module does the actual comparison of objects based on the object descriptions provided and a similarity function. An edit distance computation and comparison based on a threshold similarity measure will be done.

Edit Distance Computation: during the comparison of objects data similarity in addition to data equality needs to be covered, this helps to avoid committing an error resulted from a typographical mistake,i.e.,typographical errors need to be compensated. This consideration can be covered through an edit distance computation.

Comparison of object pairs:-after getting the object descriptions each time a pair of objects fed into the module and they are checked or compared for similarity based on thresholded similarity measure. This is based on their object descriptions and each of the corresponding object description values compared and similarity is inferred.

For the above example xml document among the object description components if at least two of the descriptions find similar with the comparison then the two objects will also be considered a match, since the total number of object descriptions is three and two of them find similar and outweighs the one object description which is not similar.

Classification Component / Module: this component classifies the objects into categories of class based on the result of the comparison function.

Classification: -based on the result of the comparison, that is duplicate pairs, a classification of the objects made into a class of matches and non-matches.

4.3.2 System A

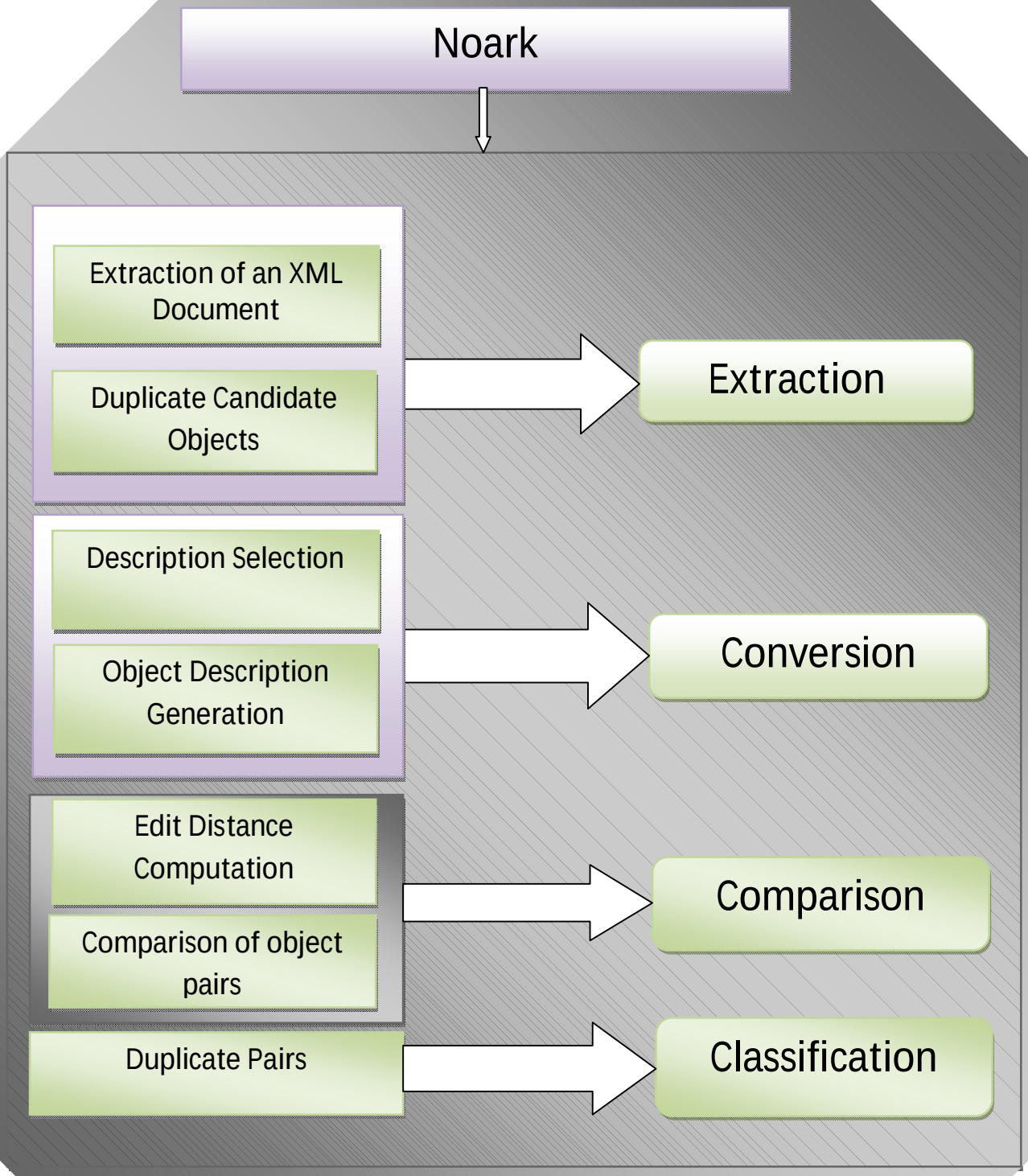
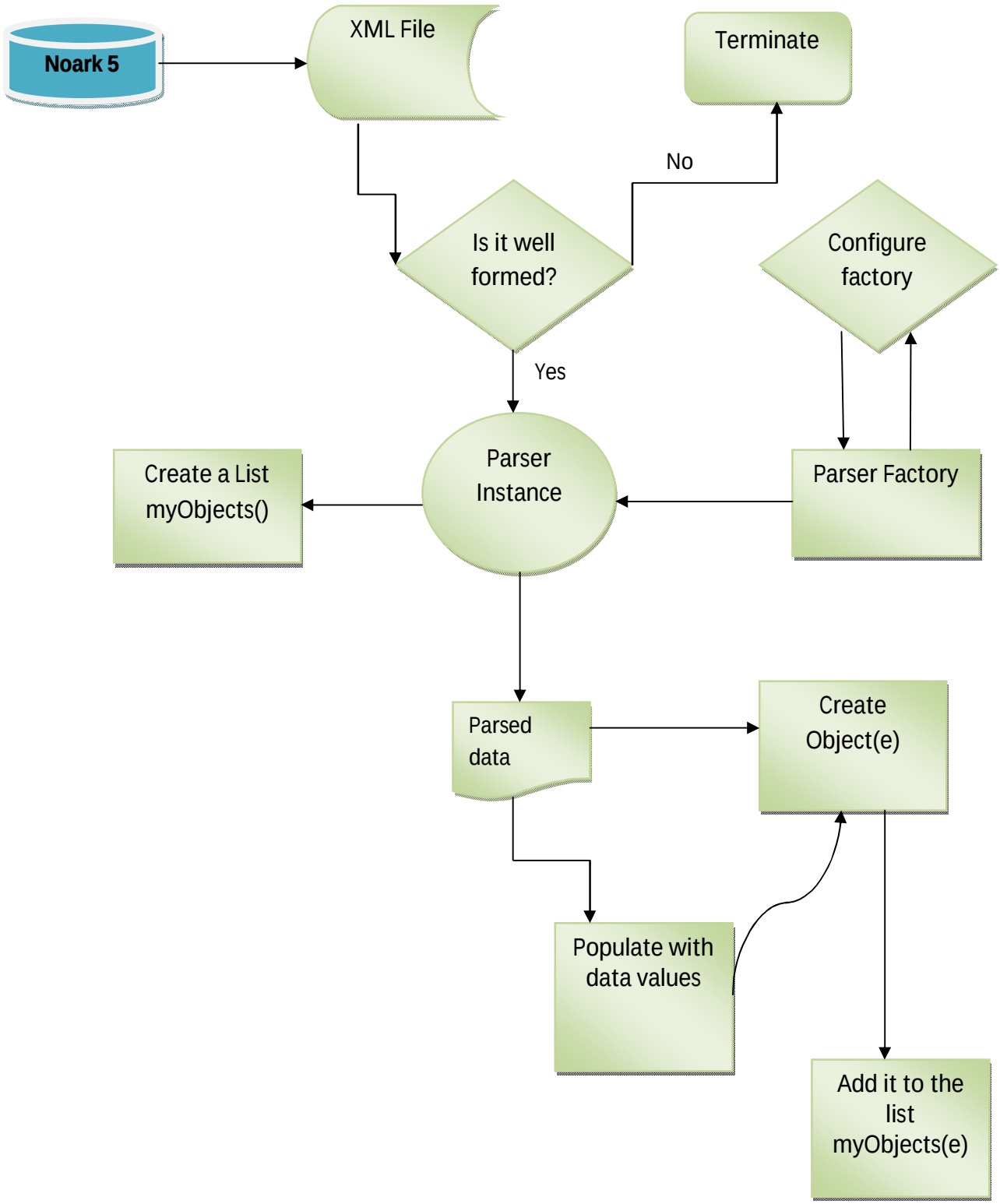


Fig.

4.3.3 Flow Chart Diagram

A flow chart is a graphical or symbolic representation of a process. Each step in the process is represented by a different symbol and contains a short description of the process step. The flow chart symbols are linked together with arrows showing the process flow direction. This diagrammatic representation can give a step-by-step solution to a given problem.

The flow chart below on fig 4.9 provides a visual representation of the functional requirements.


Noark 5
XML File
Terminate
No
Is it well formed?
Configure factory
Yes
Parser Instance
Create a List myObjects()
Parser Factory
Parsed data
Create Object(e)
Populate with data values
Add it to the list myObjects(e)

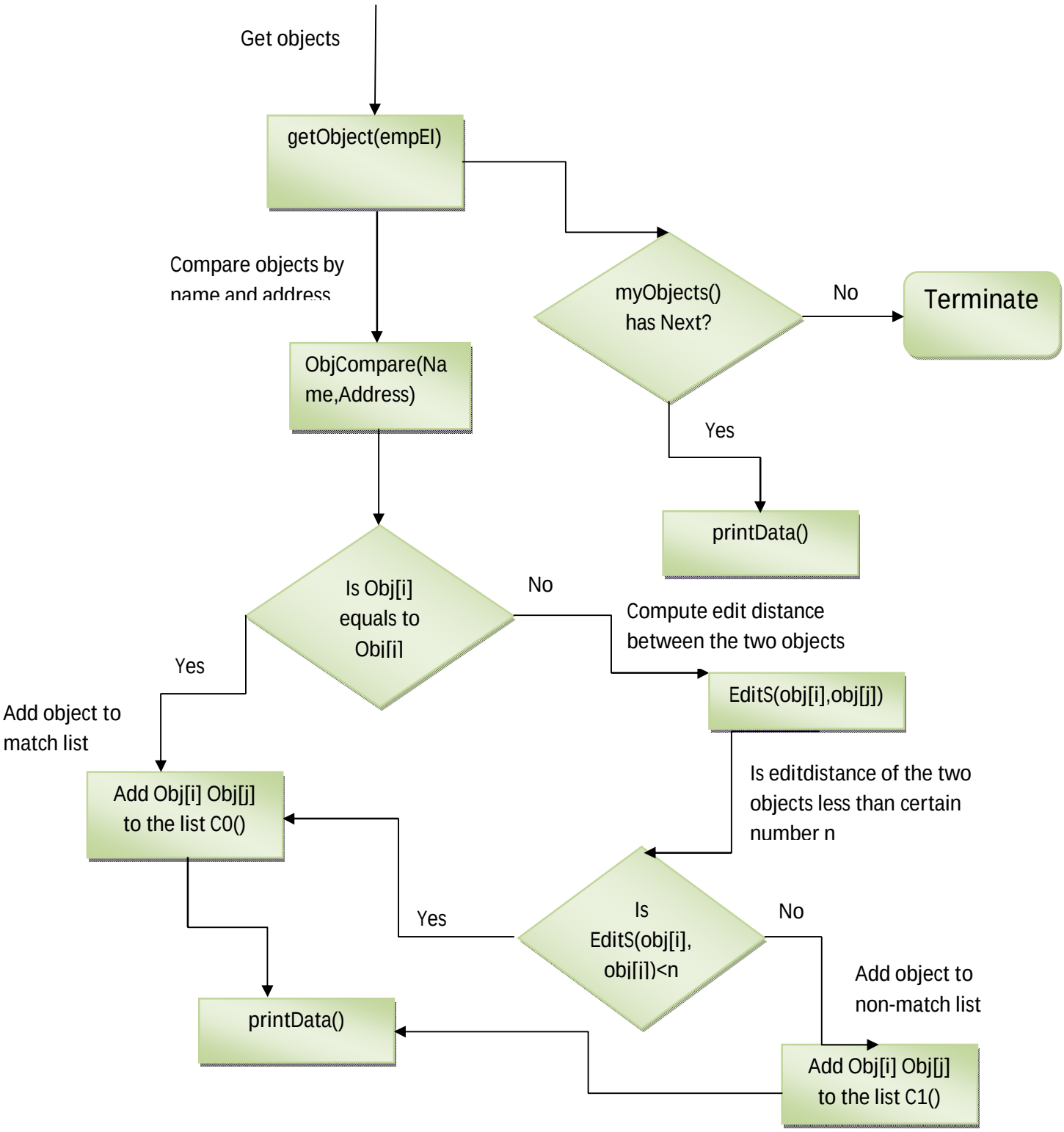


Fig.4.9 Object Identification Flow chart Diagram

4.4 Algorithm

The algorithm down shows the implementation of the object identification system, it shows the part of the solution which compares pair of objects and classifies them into a match and non-match class.

```
Object[] objA = List.myObjs()

Object[] objB = List.myObjs()

For (j=0;j<objA.length-1;j++)

  For (i=j+1;i<objB.length;i++)

    If (objA[j].equals(objB[i])) then

        Add(objA[j],objB[i]) to classmatchArrayList

    Else if !(objA[j].equals(objB[i]) then

        Compute editdistance((objA[j],objB[i]))

            If (editdisatnce((objA[j],objB[i])) < n then

              Add(objA[j],objB[i] to classmatchArrayList

            Else

              Add(objA[j],objB[i] to classnonmatchArrayList

            End if

      Else

      End if

   End for

  End for
```

Fig 4.10 Object identification technique algorithm

In the above algorithm equals method is override to consider the object description elements name and address.

Chapter -5

5.1 Experiments and Findings

This chapter includes a demonstration of the prototyped object identification technique.

The prototyped object identification technique needs to be validated by executing on a test scenario the system supposed to work, this is done to check; whether the object identification system carry out the tasks as expected,i.e.,whether the system meets the functional requirements. And it helps to identify flops or problems document them and extend the prototype.

Developing test scenarios: a test scenario is a test script that mirrors the real-world tasks that the object identification technique should be able to accomplish by running on the test. Careful planning has been given to draw up a set of test scenarios which provide broad coverage of the requirements, to determine the types of tasks the object identification technique should be able to accomplish after running on test cases, to determine whether the tasks in the scenario are realistic, complete, and representative of the tasks performed by the object identification technique prototyped. This has been met through preparing a representative data of the case study and using popular tools such as dirtyxml generator to produce artificial duplicates from clean xml data.

After a careful preparation of the test scenarios, an execution of the scenarios will be done by the prototyped object identification technique. Then an evaluation of the results will be done in order to see how the prototyped object identification technique accomplished the tasks.

Prototyping for validation

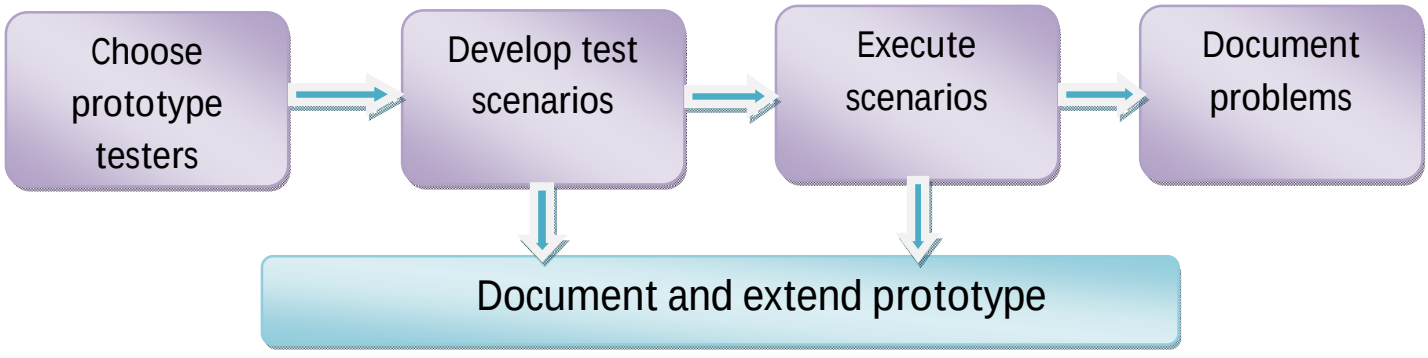


Fig 5.1 Prototyping validation process

5.1.1 Test scenario

The prototyped object identification technique in this study will take a scenario where duplicates are searched in a single XML document. And since the XML schema and XML data have been created together mostly it is unlikely that duplicates have significantly different data representation or structure. As a result errors in such type of document are mainly due to typos and missing data.

How is the algorithm handling the scenario, that means duplicates in a single xml document where errors are due to typos and missing data.

In order to handle the missing data, if the data is missing as part of a text portion within an attribute which is selected to represent in the object description then the application of an edit distance will help to handle the missing portion of text.

If the missing data is a whole text of an attribute then this will be handled by enforcing the object description selection to include only the mandatory elements from the objects attributes. In this case the object description contains the first name and last name of an employee object, as it is assumed these elements mostly find filled in databases and still they are powerful to identify a particular object.

The typographical error is also handled through a mechanism of an edit distance computation. The edit distance computation can compensate for certain typo errors by matching description values with a small number of edit distances. In this specific object identification prototype if the object description values compared and are not similar, still checked for their edit distance difference, and an edit distance of 1 is considered an acceptable range.

5.1.2 Data Sets and Setup

Dataset 1: 100 non-duplicate Employee objects created based on the Noark-5 data structure + 100 artificially generated duplicates (1 for each object)

The 100 artificially duplicated Person objects in Dataset 1 were generated automatically with an XML Dirty Data Generator.

As previously mentioned this object identification algorithm identifies duplicates where duplicates are found in a uniformly structured single xml document where there is no structural difference, the algorithm can also deal with typographical errors based on an edit distance algorithm called Levenshtein distance. So the parameters of artificially creating duplicates/contaminating a clean xml document supposed non-duplicate are set as follows:

1) Percentage of duplicates,100%
2) percentage of typographical errors 20%

Hence dataset 1 represents the scenario where mostly uniformly structured objects are duplicated by typos and missing data. Typographical errors can be achieved by the operations of Insert, delete, swap, or replace specified number of characters in text and this can easily be achieved by using the dirty XML generator as it allows setting those parameters.

An example of how artificial duplicate objects are created is given below. You can see from the figures a) persons.xml is a clean xml file b) persons_params.xml is an xml file where the parameters on how to create the duplicates can be set and c) persons_dirty.xml file is the output contaminated xml document.

Duplicate contamination Example on part of the test scenario sample data. The following sample shows a clean XML data for person object.

```
<?xml version="1.0">
<perosns>
<person ID="1" age="43" sex="m">
<firstname>Jens</firstname>
<lastname>cornor</lastname>
<Address>akersusgate.23,0173 </address>
</person>

<person ID="2" age="32" sex="f">
<firstname>Linea</firstname>
<lastname>pharo</lastname>
<Address>lakygate.13,0185 </address>
</person>

<person ID="3" age="25" sex="f">
<firstname>arni</firstname>
<lastname>sktap</lastname>
<Address>gronorgate 9,0167, </address>
</person>
</persons>
```

a) Persons.xml

The following Sample shows an XML data where the parameters of creating duplicates and contaminating those duplicates can be set.

```
<?xml version="1.0"?>

<dirtyXMLparameters
   xmlns:xs="http://www.w3.org/2001/XMLSchema-instance"
          xs:noNamespaceSchemaLocation="../dist/DirtyXMLParameters.xsd"
          valid4Desc="true"
          errorsInAncestors="false">

<algo name="swap1" baseAlgo="swapChars">
  <parameter name="includeFirstChar" value="true"/>
   <parameter name="includeLastChar" value="true"/>
   <parameter name="minSwaps" value="1"/>
   <parameter name="maxSwaps" value="2"/>
 </algo>

<algo name="swap2" baseAlgo="swapChars">
 <parameter name="includeFirstChar" value="false"/>
  <parameter name="includeLastChar" value="true"/>
 <parameter name="minSwaps" value="2"/>
 <parameter name="maxSwaps" value="4"/>
 </algo>

 <algo name="del1" baseAlgo="deleteChar">
  <parameter name="includeFirstChar" value="false"/>
   <parameter name="includeLastChar" value="true"/>
 </algo>

<algo name="ins1" baseAlgo="insertChar">
  <parameter name="includeFirstChar" value="false"/>
  <parameter name="includeLastChar" value="true"/>
 <parameter name="includeUpper" value="false"/>
 <parameter name="includeLower" value="true"/>
 <parameter name="includeDigits" value="false"/>
 </algo>

<dupElement name="person" delProb="0" dupProb="100" maxDup="1">
 <attribute name="sex">
 <chars delProb="30" changeProb="75">
 <changeAlgo algoName="swap1" useProb="100"/>
  </chars>
  </attribute>

<dupElement name="firstname" delProb="25" dupProb="60" maxDup="2">
    <chars delProb="30" changeProb="80">
   <changeAlgo algoName="del1" useProb="50"/>
```

```
   <changeAlgo algoName="ins1" useProb="50"/>
    </chars>
   </dupElement>

<dupElement name="lastname" delprob="25" dupProb="60" maxDup="2">      <chars
delProb="30" changeProb="80">
<changeAlgo algoName="del1" useProb="50"/>
 <changeAlgo algoName="ins1" useProb="50"/>
  </chars>
 </dupElement>

<dupElement name="address" delProb="0" dupProb="30" maxDup="1">
     <chars delProb="0" changeProb="70">
  <changeAlgo algoName="swap2" useProb="80"/>
   <changeAlgo algoName="del1" useProb="20"/>
   </chars>
   </dupElement>
 </dupElement>
</dirtyXMLparameters>
```

b) Persons_params.xml

This third XML data shows a duplicated version of the clean xml data created above based on the parameters set on figure b.

```
<?xml version="1.0" encoding="UTF-8"?>
<persons>
<person ID="1" age="43" sex="m">
<firstname>Jens</firstname>
<lastname>cornor</lastname>
<address>akersusgate</address>
</person>
<person ID="2" age="32" sex="f">
<firstname>Linea</firstname>
<lastname>pharo</lastname>
<address>lakygate</address>
</person>
<person ID="3" age="25" sex="f">
<firstname>arni</firstname>
<lastname>sktap</lastname>
<address>gronorgate 9,0167</address>
</person>
<person ID="1" age="43" sex="m">
<firstname>Jens</firstname>
<lastname>cornor</lastname>
<address>akersusgate.23</address>
```

<firstname>Jens</firstname><lastname>cornr</lastname></person><person ID="2" age="32" sex="f">
<firstname>Linea</firstname>
<lastname>pharo</lastname>
<address>lakygate.13,0185</address>
<firstname>Line</firstname><lastname>pharo</lastname></person><person ID="3" age="25" sex="f">
<firstname>arni</firstname>
<lastname>sktap</lastname>
<address>gronorgate 9,0167</address>
<firstname></firstname><lastname>skta</lastname></person></persons>

c) Persons_dirty.xml

Fig 5.2 Sample XML duplicate

5.1.3 Measuring the Effectiveness

Pair wise duplicate detection is evaluated using recall, precision and f-measure. They have been used extensively to evaluate duplicate detection algorithms and originate from information retrieval (Baeza-yates and Eibeiro-Neto, as cited in Weis et al., 2006). The goal is to obtain high precision for high recall resulting in a high f-measure.

Let Sall = {(o1,o'),…,(o_n,o'_n) be the set of all duplicates pairs in a data set. A duplicate detection algorithm detecting pairs of duplicates returns a set of positives:

Spos = {(oi,oi')…(oj,oj')},then the set Strue of true positives is defined as

Strue = {(oi,oi') | (oi,oi') ϵ Spos ^ ((oi,oi') ϵ Sall V (oi,oi') ϵ Sall)}

Based on the experiment among the 50 pair of candidates that means Sall the true positives returned are 38 pair of duplicates. Among them 4 pairs are found false positives. Based on the formulas given below, the recall, precision and f-measure are given in the table below.

$$\text{Recall} = \frac{|S_{true}|}{|S_{all}|}$$

$$\text{Precision} = \frac{|S_{true}|}{|S_{pos}|}$$

$$\text{F-measure} = \frac{2}{1/\text{recall} + 1/\text{precision}}$$

Non-Duplicate Employee objects	Artificially duplicated objects	Percentage of duplicates	Percentage of typographical errors	Strue(true positives)	Spos(set of positives)	Recall	Precision	F-Score
100	100	100%	20%	34	38	0.68	0.89	0.77

Table 5.1 Experiment Results of the object identification technique

Chapter-6

6.1 Conclusion

This last chapter finalizes the thesis presenting conclusion about the findings of the research and indicating what can be done in future work in the recommendation section.

This study has began by explaining how critical the data quality issue is and accepting the fact that little research has been done on an XML object identification. The study has set its goals to answer two primary questions: first developing a functional requirement of an object identification technique for Noark data and second measuring the effectiveness of the object identification technique.

In order to find the answers for the research questions the study has followed a qualitative approach focusing on a case study of the Noark (Norwegian archive), data. It has used a document analysis method to investigate the structure of Noark and develop a representative sample data for which an object identification technique has been developed. The study has also used a prototyping approach as a main methodology for developing the object identification technique as a tool, which has been a common approach in implementing information systems by helping to understand the functional requirements and developing the prototype of the object identification technique using Java Programming language tool.

What is the functional requirement of the Object identification technique? The paper has answered the question developing a functional requirement of the Object identification technique after a thorough review of the literature and investigating the Noark data structure. After understanding the functional requirements a prototype of the Object identification system has been developed and tested for validation on Noark-5 representative data. The result of the experiment shows that the study has meet the purpose, as its target was effectively identifying objects, this is a very important step for the Noark-5 as currently there is not an effort on data quality issues.

6.2 Recommendation

Despite meeting the purpose, a lot can be done to improve the object identification technique. The recommendations focus on the techniques of the Object identification:

As we have seen in this scenario the object identification technique has used mandatory elements as an object description, since to get a better result on the effectiveness and increase the accuracy of the object identification technique, we may need more object description elements which can characterize the objects, and in relation to this we have to handle missing data problem advanced way, the object description might miss some element and still needs to handle the situation.

The scenario considered on this paper reflects an Object identification within a single XML document, which mostly exhibit errors resulted due to typos and missing data since the xml schema and XML data have been created together, Extending the comparison and classification algorithm to accommodate XML data from two different sources would be a more interesting scenario, Organizations exchange different data from various sources and those data need to be quality and an object identification technique for such various types of data need to consider the structural difference of XML documents as XML data representation is flexible and could easily create problems.

It is not wise to think about effectiveness and not about the efficiency aspect, so it will be important and necessary to consider how to develop a more efficient algorithm, i.e., in this study scenario, the algorithm consider every possible pair of records for a match, and the potential number of matches is O (n2), which is large number even for datasets of moderate size. So it will be important to apply a technique which can help reduce the number of pair wise comparisons, as indicated on some studies, it will be better to have some less costy comparisons to identify a more relevant candidates prior to matching.

6.3 References

Batini, C., & Scannapieca, M. (2006). Data quality: concepts, methodologies and techniques. Berlin; New York: Springer.

W.Eckerson, W. (2002). Data Quality and the Bottom Line. In T. d. w. Institute (Ed.), Achieving Business Success through a Commitment to High Quality Data.

Universe, D. (2010). A Digital Universe decade-Are You Ready?

Agosta, L. (2008). Data Warehousing Meets Data Archiving in Information Lifecycle Management. 3. Retrieved from http://www.information-management.com/news/10001092-1.html website

Flecker, D. (2003). Digital Archiving: What is involved? Retrieved from http://net.educause.edu/ir/library/pdf/erm0316.pdf

What is Digital Preservation? (2006, 28, April, 2006). Retrieved May, 24, from http://www.digitalpreservationeurope.eu/what-is-digital-preservation/

L.Duranti. (2005). THE LONG-TERM PRESERVATION OF ACCURATE AND AUTHENTIC DIGITAL DATA: THE INTERPARES PROJECT Data Science Journal, 4(25), 12.

John Roeder, Philip Eppard, William Underwood and Tracey P. Lauriault, "Part Three—Authenticity, Reliability and Accuracy of Digital Records in the Artistic, Scientific and Governmental Sectors: Domain 2 Task Force Report," [electronic version] in International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records, Luciana Duranti and Randy Preston, eds. (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008). <http://www.interpares.org/display_file.cfm?doc=ip2_book_part_3_domain2_task_force.pdf

OCLC & CRL. (2007). Trustworthy Repositories Audit and Certification: criteria and checklist.OCLC.Retrievedfrom http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Bovee, M., Srivastava, R.P., & Mak, B. (2003). A Conceptual Framework and Belief-function Approach to Assessing Overall Information Quality. International Journal of Intelligent Systems 18(1), 51-74.

Redman, T.C. (1996), Data Quality for the Information Age, Norwood, MA: Artech House.

E.Madnick, S., Y.Wang, R., W.lee, Y., & Zhu, H. (2009). Overview and framework for data and information quality research. ACM Journal, 2, 22.

Diego Milano, Monica Scannapieco, & Catarci, T. (2006). Structure-aware XML Object Identification. Retrieved from http://pike.psu.edu/cleandb06/papers/CameraReady_122.pdf

Melanie Weis, & Naumann, F. (2005). DogmatiX tracks down duplicates in XML. 12. Retrieved from http://portal.acm.org/citation.cfm?id=1066157.1066207&coll=DL&dl=GUIDE&CFID=31054705&CFTOKEN=12935824

Thomas C.Redman. (2001). Data Quality: The Field Guide. Boston: Digital Press: Butterworth-Heinemann, Inc.

Kuan-Tsae Hunag, W.Lee, Y., & Y.Wang, R. (1999). Quality Information and Knowledge. Upper Saddle River:NY: Prentice Hall PTR.

E.Olsen, J. (2003). Data Quality: The Accuracy Dimension: Morgan Kaufmann Publishers.

Yang W.Lee, Leo L.Pipino, James D.Funk, & Y.Wang, R. (2006). Journey to Data Quality. Cambridge, Massachusetts: The MIT Press.

Redman, T.C. (1996), Data Quality for the Information Age, Norwood, MA: Artech House.

English, L. P. (1999). Improving Data Warehouse and Business Information Quality: Methods for reducing costs and Increasing Profits. New York: John Wiley and Sons,Inc.

Goasdoue, V., Nugier, S., Duquennoy, D., & Laboisse, B. An Evaluation Framework For data Quality Tools. Retrieved from http://mitiq.mit.edu/iciq/PDF/AN%20EVALUATION%20FRAMEWORK%20FOR%20DATA%20QUALITY%20TOOLS.pdf

Phillip Cykana, Alta Paul, & Stern, M. (1996). DOD Guidelines on Data Quality Management. Paper presented at the MIT Conference on Information Quality-IQ, Cambridge, A.

W.Lee, Y. (2003). Crafting Rules: Context-Reflective Data Quality Problem Solving. Journal of Management Information and Systems, 20(3), 93-119.

A. Galway, L., & H.Hanks, C. (1996). Data Quality problems In Army Logistics: Classification, Examples and Solutions: RAND.

Lee, S.-l. (2003). Data Quality in Organizational Context: A Case Study. (Master's), North Carolina, Chapel Hill, North Carolina. Retrieved from http://www.ils.unc.edu/MSpapers/2885.pdf

Elmasri, R., & Navathe, S. (2004). Fundamentals of database systems (4th ed.). Boston: Pearson/Addison Wesley.

Leo L.Pipino, W.Lee, Y., & Y.Wang, R. (2002). Data Quality Assessment. 45. Retrieved from http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf

Data Integration channels. (2011). Retrieved from Information Management website: http://www.information-management.com/channels/data_integration.html

Ravikumar, P., & W.Cohen, W. A Hierarchical Graphical Model for Record Linkage. Retrievedfrom http://www.ml.cmu.edu/current_students/ravikumar_kdd_project.pdf

Winkler, W. E. (2006), Overview of record linkage and current research directions, Technical Report RRS2006/02, US Bureau of the Census.

D.Myers, M. (1997, February 17, 2011). Qualitative Research in Information Systems. MIS Quarterly, from http://www.qual.auckland.ac.nz/

Orlikowski, W. J., & J.J., B. (1991). Studying Information Technology in Organizations: research Approaches and assumptions (Vol. 2).

Klein, H. K., & D.Myers, M. (1999). A set of Principles for conducting and evaluating interpretive filed studies in information systems. MIS Quarterly, 23(1), 67-93.

Pickard, A.J. (2007) Research methods in information.London: Facet Publishing.

Yin, R. K. (2009). In Case study research: Design and methods. Los Angeles, Calif: Sage Publications.

Verner, J. M. and N. Cerpa (1997). "Prototyping-Does your Perception Depend on Your Job?" Journal of Systems and Software, January, No. 36, pp. 3-16.

Knowledge Management made simple. Functional requirements, from http://www.ofnisystems.com/Validation/Functional_Requirements.htm

Loshin, D. (2007). Data Profiling, Data Integration and Data Quality The Pillars of Master data Management (pp. 29).

SHEILA TEJADA, CRAIG A.KNOBLOCK, & MINTON, S. (2001). LEARNING OBJECT IDENTIFICATION RULES FOR INFORMATION INTEGRATION. INFORMATION SYSTEMS, 26(8), 607-633.

WEIS, M., NAUMANN, F., & BROSY, F. (2006). A DUPLICATE DETECTION BENCHMARK FOR XML (AND RELATIONAL) DATA. RETRIEVED FROM