MEASURING THE APPLICABILITY OF OPEN DATA STANDARDS TO A SINGLE DISTRIBUTED ORGANISATION: AN APPLICATION TO THE COMESA SECRETARIAT

A dissertation submitted to the Department of Computer Science,
Faculty of Science at the University of Cape Town
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

INFORMATION TECHNOLOGY

- Themba Munalula –

Supervisor

Dr Hussein Suleman

February, 2008

ABSTRACT

Open data standardization has many known benefits, including the availability of tools for standard encoding formats, interoperability among systems and long term preservation of data. Mark-up languages and their use on the World Wide Web have implied further ease for data sharing. The Extensible Markup Language (XML), in particular, has succeeded due to its simplicity and ease of use. Its primary purpose is to facilitate the sharing of data across different information systems, particularly systems connected via the Internet.

Whether open and standardized or not, organizations generate data daily. Offline exchange of documents and data is undertaken using existing formats that are typically defined by the organizations that generate the data in the documents. With the Internet, the realization of data exchange has had a direct implication on the need for interoperability and comparability. As much as standardization is the accepted approach for online data exchange, little is understood about how a specific organization's data "*fits*" a given data standard. This dissertation develops data metrics that represent the extent to which data standards can be applied to an organization's data.

The research identified key issues that affect data interoperability or the feasibility of a move towards interoperability. This research tested the unwritten rule that organizational setups tend to regard and design data requirements more from internal needs than interoperability needs. Essentially, by generating metrics that affect a number of data attributes, the research quantified the extent of the gap that exists between organizational data and data standards. Key data attributes, i.e. completeness, concise representation, relevance and complexity, were selected and used as the basis for metric generation. Additional to the generation of attribute-based metrics, hybrid metrics representing a measure of the "goodness of fit" of the source data to standard data were generated.

Regarding the completeness attribute, it was found that most Common Market for Eastern and Southern Africa (COMESA) head office data clusters had lower than desired metrics to match the gap highlighted above. The same applied to the concise representation attribute. Most data clusters had more concise representation for the COMESA data than the data standard. The complexity metrics generated confirmed the fact that the number of data elements is a key determinant in any

move towards the adoption of data standards. This fact was also borne out by the magnitude of the hybrid metrics which to some extent depended on the complexity metrics.

An additional contribution of the research was the inclusion of expert users' weights to the data elements and recalculation of all metrics. A comparison with the unweighted metrics yielded a mixed picture. Among the completeness metrics and for the data retention rate in particular, increases were recorded for data clusters for which greater weight was allocated to mapped elements than to those that were not mapped. The same applied to the relative elements ratio. The complexity metrics showed general declines when user-weighted elements were used in the computation as opposed to the unweighted elements. This again was due to the fact that these metrics are dependent on the number of elements. Hence for the former case, the weights were evenly distributed while for the latter case, some elements were given lower weights by the expert users, hence leading to an overall decline in the metric.

A number of implications emerged for COMESA. COMESA would have to determine the extent to which its source data rely on data sources for which international standards are being promoted. Secondly, an inventory of users and collectors of the data COMESA uses is necessary in order to determine who would be the beneficiary of a standards-based information system. Thirdly, and from an organizational perspective, COMESA needs to designate a team to guide the process of creation of such a standards-based information system. Lastly there is need for involvement in consortia that are responsible for these data standards. This has an implication on organizational resources.

In totality, this research provided a methodology for determination of the feasibility of a move towards standardization and hence makes it possible to answer the critical first stage questions such a move begs answers to.

ACKNOWLEDGEMENTS

Thanks are due to my supervisor, Dr Hussein Suleman, for his patience and guidance throughout the research. Thanks also are due to my work colleagues who participated in the user evaluation. Last but not least thanks are due to my family for their understanding and patience.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Motivation

Open data standardization has many known benefits, including the availability of tools for standard encoding formats, interoperability among systems and long term preservation of data. Markup languages and their use on the World Wide Web have implied further ease for data sharing. The Extensible Markup Language (XML), in particular, has succeeded due to its simplicity and ease of use. Its primary purpose is to facilitate the sharing of data across different information systems, particularly systems connected via the Internet.

Whether open and standardized or not, organizations generate data daily. Offline exchange of documents and data is undertaken using existing formats that are typically defined by the organizations that generate the data in the documents. With the Internet, the realization of data exchange has had a direct implication on the need for interoperability and comparability. As much as standardization is the accepted approach for online data exchange, little is understood about how a specific organization's data "*fits*" a data standard.

An organization can be seen as a single distributed entity that generates data daily through its various transactions and workflows. This could include typical data such as financial data and specialized data such as economic statistics. Financial data will typically follow an organization-specific format, though a number of the generated fields have a generic application e.g., total cost in an order document may be expected in the data of other organizations.

The more specialized data such as economic statistics may not easily conform to the data formats of other organizations.

With regard to the two data clusters above, data standards exist and are designed by a community of users. For financial data, for instance, we have the Business Application Software Developers (BASDA) standards (BASDA, 2004) while for economic statistics we have the Statistical Data and Metadata Exchange (SDMX) standards (SDMX, 2005).

With the increasing popularity of XML-based data formats, it is possible to represent many granular and aggregate data entities in standards-based XML. While the existence of data standards is acknowledged, generation of metrics to assess the feasibility of migration is determined by the organization in question and indeed the type of data. This research seeks to use the methodology of metric generation as a way of assessing a move to standardization for an organization.

1.2 About COMESA Secretariat

COMESA is a regional integration organization dedicated to facilitation of trade and investment in the Eastern and Southern African region. Its current membership is 20 countries ranging from Libya in North Africa to Swaziland in Southern Africa. This study used data from COMESA and focused on human resource, finance and statistical data that was available for the research. Typically the human resource data is used in recruitment and staff placement issues while the finance data used in the research is used for payments and any interaction with suppliers. Statistical data is used in trade policy research.

1.3 Problem Statement

Is a move to standardization feasible? We seek to understand the degree to which standards can be applied to COMESA Secretariat data through the generation and analysis of relevant metrics.

The metrics will contribute to an understanding of the organizational data in relation to an existing data standard. This study looks at how XML-based data standards may be applied to the data storage and workflow needs of a single distributed organization.

Ultimately, we seek to provide evidence of the extent to which any possible move towards greater standardization will affect operations of the organization.

1.4 Experimental Approach

This section briefly highlights the approach we adopt in the research. First COMESA organizational data flows are analyzed. Secondly, using XML schemas from existing data standards, we map COMESA data to these standards and using a tool we designed, metrics are generated for how well the data fit the standards and vice versa. The metrics are generated based on a relevant measurement framework by Kaner and Bond (2004).

Secondly we define application profiles in order to adapt to the functional requirements of the COMESA data sets while retaining interoperability with the target standard.

An evaluation of the tool is undertaken as the final stage of the research.

1.5 Dissertation Outline

The report will be divided into six chapters. Chapter 1 is the introduction; chapter 2 is the background and literature review; chapter 3 discusses the methodology; chapter 4 is the analysis of results; chapter 5 is the evaluation; and, finally, chapter 6 is the conclusion.

CHAPTER 2
BACKGROUND

2.1 Interoperability

2.1.1 The Significance of Interoperability

"Interoperability needs more attention than ever" [European Information and Communication Technology Industry Association(EICTA), 2004]

The digital format of information and connectivity of user devices to many sources of content has resulted in the need for interoperability in these services and devices. It has been argued that while much discussion on interoperability has been technical, the attainment of interoperability should ultimately be measured by a user's experience [European Information and Communication Technology Industry Association, *op. cit.*]. Hence, ultimately, interoperability is achieved when this expectation of the user to exchange and use information among various devices and software and from many service providers is met.

The EICTA White Paper further argues that the only technical barriers to interoperability should come from limitations of technology and not those introduced for the purpose of removing interoperability by vendors and service providers.

Linked to this argument is the need for open standards to promote and focus on those elements that are necessary for the fulfillment of the interoperability requirement. This approach allows room for innovations or additions outside the scope of interoperability.

Interoperability is significant for a number of reasons:

- For content and service providers – Gives them the ability to reach a maximum audience.
- For developers – Affords predictability that software programs will work on a maximum number of environments.

- Vendors of servers, networks and client solutions – Affords these an unfragmented global market.
- Users – User experience is enhanced by the convenience of faster and better information flow as well as heterogeneous multi-vendor solutions that work seamlessly without external intervention.

As the exact opposite of fragmentation, interoperability therefore has a major impact at the macro and micro levels of an economy given its impact on competitiveness and efficiency.

2.1.2 What is Interoperability

The term "interoperability" is defined as:

"The ability of two or more systems or components to exchange information and to use the information that has been exchanged" (IEEE, 1990)

Information that is exchangeable in the definition of interoperability can be of any kind, such as voice, pictures, documents and software code.

As such, interoperability manifests itself in user satisfaction. The absence of user intervention in the exchange of information between different platforms, networks, applications or devices is the expected deliverable of interoperability. It can be deduced that the inevitability of variations in user experience underpins the importance of interoperability to deliver in accordance with this user paradigm.

2.1.3 How do you achieve Interoperability?

Many aspects of our daily lives depend on standards. Standards influence the products we use, the foods we eat, how we communicate, our trade, our means of travel, our modes of work and play, and many other activities (for example World Heath Organization, 2005, International Telecommunications Union, 2007 and Open Travel Alliance, 2007).

They may function to inform, facilitate, control or indeed a combination of such elements. They serve economic ends, by enabling or facilitating commercial transactions. They also have a social objective, such as protecting health, safety, and the environment.

Using the International Organization for Standardization (ISO) definition a standard is a:
"Document established by consensus and approved by a recognized body, that provides for common and repeated use, rules, guidelines or characteristics for activities or their results aimed at the achievement of the optimum degree of order in a given context" (ISO, 1996)

As discussed in the following sections, interoperability is achieved in the following ways:

2.1.3.1 Standard Interfaces

A standard interface describes certain generic requirements that a technical implementation of that interface needs to match in order to produce the desired functionality. Related to data interoperability two areas of information are referred to:

- Data formats – This is the particular way that information is packed into the digital package in order to allow for successful unpacking and reading of the information, according to the defined description of the information. Examples of data formats are compression file formats such as ZIP, JPEG, GIF and MPEG as well as markup language data formats such as DocBook, HTML and XML. The Open Directory has compiled 1,956 data formats (Open Directory Project, 2007a).
- Protocols – This is the sequence and meaning of information in the various data packages transferred between two interoperable elements. Examples of these are hypertext transfer protocols, domain name system and file transfer protocols. Close to 838 protocols related to the internet are available on the Open Directory website (Open Directory Project, 2007b).

A critical characteristic of these interfaces is that they be open. According to the EICTA (2004), a standard is open if it conforms to the following four criteria:

- Control – The change in the specification of the standard is done in a transparent process that is open to all interested contributors.
- Completeness – The technical requirements of the standard should be specified completely.
- Compliance – This relates to the adoption of the standard by a broad group of implementers such that interoperability is achieved via the wide availability of implementations.

- Cost – Fair and non-discriminatory access is provided to the intellectual property used in the implementation of the standard. This ensures the standards are available to interested parties at no charge or at a nominal charge.

2.1.3.2 Overview of Open Data Standards

Open data standards are standards that ensure interoperability between different solutions that need to operate on or use the same data. These standards are designed to describe data formats such that anyone can read, write or update data using tools that suit their present needs. Further they are often written by *independent* organizations and are publicly documented and freely available.

Examples of open data standards are:

- Text based standards such as
  - Extensible MarkUp Language (XML) - XML is a general purpose markup language (World Wide Web Consortium, 2006). XML's primary purpose is to enable the sharing of structured data across different information systems. It allows users the flexibility to define their own elements.
  - Resource Description Framework (RDF) –The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise and community boundaries (World Wide Web Consortium, 2004). It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on RDF, which is used to represent information and to exchange knowledge in the Web. W3C's role is in drawing attention to the RDF specification and promoting its widespread deployment. This enhances the functionality and interoperability of the Web.
  - HyperText MarkUp Language (HTML) – HTML is a markup language that describes the structure of text-based information in a document. HTML uses tags such as <h1> and </h1> to structure headings, paragraphs, etc. (World Wide Web Consortium, 1999). It can also describe, to some extent, the appearance and semantics of a document.
  - Electronic Business using eXtensible Markup Language (ebXML) is a modular suite of specifications that enable companies of any size and in any location to undertake business over the Internet (OASIS, 2002). Through its use, companies now have a standard

method to exchange business messages, conduct trading relationships, communicate data in common terms and define and register business processes.

- DocBook - DocBook is a schema used to describe books which is available in several languages including XML DTDs and W3C XML Schema (OASIS, 2006). It is maintained by the DocBook Technical Committee of OASIS and has been adopted by a large and growing community of authors writing books of all kinds. DocBook is supported by a number of commercial tools, and a growing number of free software environments, which qualifies it as an open standard.

- Binary formats such as
  - The Joint Photographic Experts Group (JPEG) Standard deals with issues pertaining to the discussion and creation of standards for still image compression (International Telecommunications Union, 1993).
  - Portable Network Graphics (PNG) is a format for storing bitmapped (raster) images on computers (Roelofs, 2007). PNG was developed in answer to the GIF format, which became decidedly less useful when Unisys and CompuServe suddenly announced that programs implementing GIF would require royalties because of Unisys' patent on the LZW compression method used in GIF (See Roelofs, *op. cit*)). This announcement only catalyzed the development of a new and much-improved open replacement format. PNG is the result.

2.1.3.3 Proprietary Specifications

In this instance a party owns or exercises control over the specification and its use. These proprietary specifications can contribute to interoperability, particularly at the introduction of new solutions. However there are no guarantees that a proprietary product can or will become more open.

2.1.3.4 Open Source Software

Open source software is software whose source code is published and available publicly, thus allowing anyone to copy, modify and redistribute the code without payment of fees or royalties (See Open Source Initiative, 2006). Whereas the concept of open source is distinct from open standards, it is largely so because it is regarded as an implementation while a standard is a specification.

However, contingent on certain factors, open source implementations do promote interoperability. We highlight those factors below:

- Where open source licensing allows for distribution and usage of software without any restrictions, the resulting network effect has the capacity to promote standard usage and in this way contribute to better interoperability.
- The fact that open source software is transparent means it promotes trust in its interoperability.
- The platform portability characteristic of most open source software means it can support wider dissemination which lends itself to supporting interoperable implementations.

2.2 XML AND XML STANDARDS

2.2.1 Extensible Markup Language (XML)

XML is a markup language for documents containing structured information. Names, allowable hierarchy and element data types and attributes are defined in an XML schema. XML schemas utilize a rich data typing system that allows for detailed constraints on an XML document's logical structure (World Wide Web Consortium, 2006).

2.2.2 XML-Based Data Standards

XML-based data standards are widely used. Many industry consortia have issued design guidelines and patterns for developing libraries of schemas in accordance with the respective profiles. Examples of such consortia, with the first three being the ones that will be used for evaluation of COMESA data, are:

- The Human Resource–XML (HR-XML) standard deals with the development and promotion of a standard suite of XML specifications that enable the exchange of human resource information (HR-XML Consortium, 2007). The HR-XML consortium is responsible for the development of these standards. Its library deals with the following areas of HR: recruitment, competencies, assessments, performance management, background checks, payroll, employee benefits, staffing and metrics.
- eBIS-XML is the Business Application Software Developers electronic business interchange standard in XML(BASDA, 2004) . Among the XML schemas provided to define the

standard are: Order, Order Response and Invoice. BASDA is an association representing 200 leading software suppliers.

- Statistical Data and Metadata Exchange standards facilitate the exchange of statistical information (SDMX, 2005). Time series as well as cross sectional XML formats are supported. The SDMX initiative is sponsored by several international organizations involved in statistics
- Dublin Core is a baseline metadata standard for electronic resources developed by the Dublin Core Metadata Initiative (Dublin Core Metadata Initiative, 2002).
- Chem eStandards are standards developed for the purchase, selling and delivery of chemical products. The Chemical Industry Data Exchange (CIDX) is a non profit trade association behind these standards (CIDX, 2004). They are XML-based and are the cooperative effort of chemical companies.

2.3 Data Quality Metrics

Data interoperability is regarded as a pervasive, long term and expensive problem. Hughes (2006) suggests that as an issue, research into data interoperability has not received support commensurate with the severity of the problem. Assessing how suitable a single organization's data is in relation to existing data standards and vice versa requires relevant data metrics. Given the multidimensional aspect of data quality, the development of metrics has both subjective and objective viewpoints. These criteria largely depend on the needs and experiences of the users of the data.

Pipino *et al.* (2002) propose a set of principles that assist organizations to develop a set of usable data quality metrics. Table 1 below summarizes the dimensions of data quality that they discuss.

Table 1: Data Quality Dimensions

Dimensions	Definitions
Accessibility	The extent to which data is available, or easily and quickly retrievable
Appropriate Amount of Data	The extent to which the volume of data is appropriate for the task at hand
Believability	The extent to which data is regarded as true and credible
Completeness	The extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	The extent to which source data is presented in the same format as destination data
Ease of Manipulation	The extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	The extent to which data is correct and reliable
Interpretability	The extent to which data is in appropriate languages, symbols and units, and the definitions are clear
Objectivity	The extent to which data is unbiased, unprejudiced and impartial
Relevancy	The extent to which data is applicable and helpful to the task at hand
Reputation	The extent to which data is highly regarded in terms of its source or content
Security	The extent to which access to data is restricted appropriately to maintain its security
Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand
Understandability	The extent to which data is easily comprehended
Value-Added	The extent to which data is beneficial and provides advantages from its use

Source: Pipino et al.

Pipino *et. al.* (*op cit)* present three functional forms that assist in developing data quality metrics. These forms are:

- Simple ratio: which measures the ratio of the desired outcome to total outcomes.
- Min or Max Operation – This form relates to dimensions that require aggregation of multiple data quality indicators.
- Weighted Average – In the multivariate case, this forms the alternative to the min operator. If there is a good understanding of the importance of each variable to the overall assessment of a dimension, then a weighted average of variables is appropriate.

They further present an approach that combines both subjective and objective criteria. They define a four quadrant matrix with increasing objectivity on the x-axis and increasing subjectivity on the y-axis as illustrated in Table 2.

Table 2: Quadrants of Data Quality Assessment

II – High subjectivity & low objectivity	IV - High subjectivity & high objectivity
I- Low subjectivity & low objectivity	III - Low subjectivity & high objectivity

Applying this hybrid criterion to two firms, they find that one firm falls within the scope of what they call quadrant I, while the other falls within the scope of quadrant IV. An important conclusion they make is the fact that a "one size fits all" set of metrics are never the solution. They suggest that the assessment of data quality is a continual effort that requires the awareness of fundamental principles underlying the development of both subjective and objective data quality metrics.

Loshin (2005) provides insight into how organizations can develop data quality metrics in conjunction with their clients. He does this by defining 8 principles for defining such a metric: clarity of definition, measurability, business relevance, controllability, representation, reportability, tractability and drill down capability. He proceeds to divide organizational data into two categories: those that impact the achievement of business operational and strategic goals and those that do not. For each perceived problem, the process is broken down into:

- A review of how the data flaw relates to each area of impact.
- Determination of the frequency with which impact is incurred.
- Summing up the measurable costs associated with each impact experience by the data quality problem.
- Assignment of an average cost to each occurrence of the problem.

Using the business relevance areas, namely profitability, productivity, risk and intangibles, he highlights how the above four bullet points can be applied to these. He concludes by suggesting the inclusion of the reporting and presentation of these metrics by the data analyst to the business client impacts the achievement of business objectives and consequently determines "hard costs" associated with each occurrence of a flaw.

The IEEE Standard 1061(IEEE, 1998) lays out a methodology for the development of software quality metrics. Concerning the development of the metrics, it lays out the following points as validation criteria:

- Correlation – A metric should be linearly related to the quality factor.
- Consistency – Let Q be the quality factor and X the output of the metrics function, H:Q->X. H must be a monotonic function i.e. $q_1>q_2>q_3$ implies $x_1>x_2>x_3$.
- Tracking – Given the functions above, as Q changes from q_1 to q_2, X changes promptly from x_1 to x_2.
- Predictability – If the value of Q is known at some point in time, then we should be able to predict the value of X.
- Discriminative power – A metric should be able to discriminate between high quality components and low quality components.
- Reliability – a metric should demonstrate correlation, tracking, consistency, predictability and discriminative power properties for at least D% of its application, D% being a pre-agreed threshold for reliability.

Kaner and Bond (2004) propose a framework for evaluating metrics. Their framework essentially captures the essence of *construct validity* i.e. asking the question: *How do you know that you are measuring what you think you are measuring*? The framework goes on to ask a set of critical questions:

- What is the purpose of this measure?
- What is the scope of this measure?
- What attribute are we trying to measure?
- What is the natural scale of the attribute we are trying to measure?
- What is the natural variability of the attribute (defined as a measurable physical or abstract property of an entity)?
- What is the metric or function that assigns a value o the attribute?
- What is the natural scale of this metric?
- What is the natural variability of readings from this instrument?
- What is the relationship of the attribute to the metric value?
- What are the natural and foreseeable side effects of using this instrument?

Kaner and Bond (*op. cit.*) proceed to apply this framework to bug count metrics. They conclude that commencing from a detailed analysis of the task or attribute under study might lead to more complex metrics but this effort also leads to more meaningful and therefore more useful data.

Ochoa and Duval (2005) address the issue of low quality metadata through a quality evaluation framework that incorporates users' or human quality review. They find that one of their proposed metrics, text information content, when applied to sample records obtained from a repository is a good predictor of the human quality evaluation.

2.4 Summary

This chapter has provided a background to the data interoperability issue and matters arising in approaching its measurement. The chapter highlighted the significance of interoperability as well as defining it and explaining how it is achieved. With regards to data standards the chapter provided an overview of existing XML based data standards and discussed the issue of measurement of interoperability of organizational data with such standards using data metrics.

What is evident when one considers the principles of metric design in measuring data interoperability is the fact that no one size fits all. This flows into our next chapter where we follow a systematic approach in determining suitable metrics for our research. The approach largely uses Kaner and Bond's (*op. cit*.) holistic approach to determining what we are trying to measure. Following Ochoa and Duval (2005) we propose, where possible, to incorporate a human evaluation aspect to the measurement of some of the metrics we propose.

CHAPTER 3

METHODOLOGY

This chapter firstly defines the principles that are adopted in pursuit of producing metrics that yield an understanding of the comparability of COMESA data against existing data standards. Secondly it presents the methodological flow and ends with the definition of the actual metrics.

3.1 Application of the Kaner and Bond Framework to COMESA Data

Critical to the Kaner and Bond framework is the definition of metrics not so much in terms of its functions, but instead in terms of the question that seeks to be answered. As such the nature of the information or attributes assist in answering this question. Hence the framework as applied to this study looked at particular attributes of data metrics for assessing the move to standardization.

In this section the Kaner and Bond framework discussed in chapter 2 is applied to COMESA data.

a. What is the purpose of this measure?
   - Evaluation of how well COMESA data elements map onto data standard elements.
   - Evaluate the implications on COMESA's information systems of adoption of the data standards.

b. What is the scope of these measures?

These metrics will be used within COMESA as an organization.

c. What attributes are we trying to measure?
   - Completeness
   - Relevance
   - Concise representation
   - Complexity (or understandability according to Table 1 of the literature review section)

d. What is the natural scale of the attribute we are trying to measure?

This is not necessarily intuitive. A natural scale does not easily come to mind for all four attributes above. Hence, a priori, no knowledge of the natural scale of these attributes exists.

e. What is the natural variability of the attribute?

For any metric that is adopted to represent an attribute such as completeness, the measure may differ depending on factors such as users' subjective opinions. Hence, the inherent sources of variation may be due to:

- Users' subjective judgments in deciding on weights.
- The fact that a user excludes certain elements as repeat fields while another user includes them. This perhaps reflects the fact that as far as interpretation of the role of elements in standards or indeed in organizational data, the element counts, etc., that are fundamental to measuring some of the metrics in this paper are not perfectly deterministic. This becomes a source of variation.

  Hence, a priori knowledge of the natural variability of the above attributes does not exist.

f. What is the natural scale of these metrics

The natural scale for these metrics is interval, ratio or ordinal. For the metrics under consideration in this study, the natural scales of the metrics are between 1-100%, 0-1 and 0-∞. Table 3 below gives the scales for the proposed metrics in this study.

Table 3: Metrics and their Scales

Attribute and Metrics	**Natural Scale**
Completeness	
Number of elements, number of source elements, number of mapped source fields, number of mapped destination fields, multiple mapping fields, relative element counts	0-∞
Data Retention Rate, destination redundancy rate	0-1
Relevance	
Aliases, synonyms, Homonyms	0-∞
Alias redundancy factor, synonym redundancy factor, homonym redundancy factor	0-1
Concise Representation	
Information to markup ratio	0-1
Complexity	
Mapping additions ratio	0-1
Structured document complexity metric	0-∞

g. What is the relationship of the attribute to the metric value?

It is clear that the attributes of data quality that are being measured do not have generally agreed methods of measurement. Hence the use of surrogate measures is adopted whereby numbers are unambiguously assigned according to rules. These details are captured in section 3.2.3 below. Table 4 below shows the relationships between the metric value and the attribute value.

Table 4: Relating Value of Attributes to Value of Metrics

Attribute and Metrics	**Relationship between value of attribute and value of metric**
Number of elements, number of source elements, number of mapped source fields, number of mapped destination fields, multiple mapping fields, relative element counts	The larger the number of mapped elements the more complete the data source schema. The fewer the elements that are mapped to the same destination element the less the redundancy in the source data schema. The higher the relative elements ratio, the more complete the source data schema.
Data Retention Rate(DRR), Destination Redundancy Rate(DeRR),	The higher the DRR, the more complete the source data schema. The lower the DeRR is, the less redundant the source data schema.
Aliases, synonyms, Homonyms	The more elements are aliases and synonyms the less relevant these source data schema elements are.
Alias redundancy factor, synonym redundancy factor, homonym redundancy factor	The higher the alias and synonym redundancy factors the more redundant the source data elements are.
Information to markup ratio	The higher this ratio is, the more concise our information.
Mapping additions ratio	The more diverged this ratio is from 1, the more extra capability is being added that is not required for mapping purposes.
Structured document complexity metric	A higher SDC metric implies more complexity.

3.2 Methodological Flow

3.2.1 Data Standards Schema Survey

A survey of existing data schemas and their sources was undertaken. Each of the schemas is formally described using XML schema and is defined in chapter 4.

Data schemas that are applicable to the COMESA data were thereafter selected.

3.2.2 Schema level transformations

The issue of data transformations is a challenging one. In the database context, a data transformation is represented by an expression $Z_i = f(Z_j)$, where Z_j is an instance of schema S_j of the source data and Z_i is an instance of the schema S_i of the target database.

Okawara *et. al.*(2005) state that it is generally regarded as difficult to develop a function *f*, from information on schemas Z_i and Z_j only. In fact their description of a tool they use to match metadata requires that the schema components are extracted in advance from both the target and source schemas. It is proposed to adopt a similar approach of extraction of elements from both the source and destination prior to the mapping.

The data standard schemas utilized requires that some prior data cleaning is done then transformations of records will be undertaken.

The process of mapping links source data elements to destination schema elements that are similar in meaning. This process is in fact at the core of this experiment. Details of this process are further elucidated in chapter 4 on design and implementation.

3.2.3 Description of the metrics used in our evaluation

The logical starting point in analyzing data is to identify the basic work unit called a file (XML document, in the case of our study). Linked to the XML document is an XML schema. The schema contains a complete representation of elements. The elements in the relevant data schemas are compared to those in the COMESA data. It must be noted that Table 1 (in the literature review section) representing dimensions and definitions does not always lend itself to data uniquely having one dimension. It may be that several dimensions are represented in one metric.

The metrics proposed below are computed together with an alternate set of metrics that have included in their computation user weights. A weighting profile ranging from 1-10 was proposed to

the expert users. Each user was provided a weight for each element. Appendix 1 presents these weights for each element. This average was used for the computation of the weighted metrics. The following example illustrates our use of user weights in the computation of weighted metrics for the data retention rate (DRR). The un-weighted metric as shown below is the ratio of the number of mapped source fields to the total number of source fields. Hypothetically suppose we have a total of six elements and three are mapped. Users give a weight for each on a scale of 0-1 based on their perceived importance in their specific subject area. Assume further that the three mapped elements receive weights of 1, 0.5 and 0.6 respectively while the unmapped elements receive weights of 1, 0.7 and 1 respectively. In order to compute both the weighted and unweighted DRR, we first compute the total number of source elements in each case. For the unweighted case this is clearly 6. For the weighted case, we add the weights to obtain a weighted total of 4.8. The effect of the weighting demonstrates the perceived importance of that element for the user and therefore affects the total weighted element count. Secondly we need to determine the weighted total for the mapped elements, which gives 3 for the unweighted case and 2.1 for the weighted case. Hence the weighted DRR is given as 2.1/4.8 or 0.44 compared to 3/6 or 0.5 for the un-weighted case. Table 5 below summaries these computations.

Table 5: Weighted and unweighted metric computation illustration

Mapping Status	Elements	Unweighted Count	Weighted Count
Mapped	A	1	1
Mapped	B	1	0.5
Mapped	C	1	0.6
Unmapped	D	1	0.7
Unmapped	E	1	1
Unmapped	F	1	1
Summary	Total # source elements =A1	6	4.8
	Total # mapped elements = B	3	2.1
	DRR=B/A1	0.5	0.44

- **Completeness Metrics**

Measured completeness metrics

Comsys (2002) provides an extensive list of data definition metrics. This section uses some of these metrics to define the attributes discussed in section 2.3..

- Number of total elements (A) – This is a count of all defined element names in the destination schema. Copy or repeat element names should not be included. Jelliffe (2006a) also proposes simple element counts as a beginning to development of data metrics. A weighted total number of destination field elements, Á, is also computed. These are weighted by users and an average of users' weighting obtained to give Á.
- Number of total source elements (A1) - Count of all defined elements in source data.
- Number of source fields mapped (B) – This is the number of source fields that are mapped. Data elements are atomic units of data that actually have a data element name and definition and might have an optional enumerated value of code. Note that it is the identification of a relationship between the data element names in the source and destination schemas that we refer to as mappings.
- Number of destination fields mapped (C) – This is the number of destination fields that are mapped.
- Multiple mapping (M) – elements in the destination format that are mapped to the same data element in the source format.

Derived Completeness Metrics

- Jelliffe(2006b) proposes a *data retention rate(DRR) or mapping completeness ratio* as a metric for assessment of data completeness. This is given by B/A1 in our data. With this metric we are able to tell how much of the original data has been mapped.
- *Destination redundancy rate or DeRR* – M/B. With this metric, one is able to represent the degree of redundancy in our mapping.
- *Relative Element Counts* –A/A1. This metric represents the ratio of the elements in the source schema to those in the COMESA data.

- *Weighted Completeness Metrics* – Here the relative element counts and destination redundancy rates are re-computed using Á in place of A. Inherent in these calculations is a reflection of users' assessment of the data standard.

- **Relevance Metrics:**

  The Comsys(2002) also provides a list of data metrics that measure the relevance attribute. The following metrics are computed for the destination format in order to establish redundancy in the data for a given mapping.

  Measured Relevance Metrics

  - Number of aliases (D) – These are redundant data elements representing the same physical data with different root data names.
  - Number of synonyms (E) – These are similar to aliases but differ in that they use different definitions for the data elements.
  - Number of homonyms (F) – These are data elements with the same name but representing different physical data.

  Derived Relevance Metrics

  - Alias redundancy factor – D/A
  - Synonym redundancy factor – E/A
  - Homonym redundancy factor – F/A
  - *Weighted relevance metrics* – Here the alias, synonym and homonym redundancy factors are recomputed using Á in place of A. Inherent in these calculations is a reflection of users' assessment of the data standard relevance attributes.

- **Concise representation Metrics**

The following metric represents the dimension of concise representation.

- Information to markup ratio. This ratio indicates how concise the representation of our data is. It will be computed as a ratio of the number of bytes of information to the number of bytes of markup. The higher the ratio the more concise is our information.

- **Complexity Metrics**
  - **Mappings Additions Ratio**

    Jelliffe(2006c) proposes a metric that asks the question "How many fields are in the intended schema that are not in the original schema?". This metric reflects the

consequence of addition of extra elements to support from the organization's perspective. It is expressed as *A1/A*, and is known as the *mapping additions ratio or MAR*. The further it diverges from 1, the more extra capability is being added that is not required for mere mapping purposes. If extra capability and its resultant costs do not correspond to business requirements, the destination schema should be pruned if possible or declared unsuitable.

- **Structured Document Complexity (SDC) Metric**

  This metric seeks to find out how complex the schema or document is in order to understand the implications from a project implementation perspective. Jelliffe(2006d) proposes a complexity metric that involves addition of the total number of elements, unique attributes, an extra point for every mandatory element, an extra point for every mandatory attribute and an extra point for every element that can only appear in position 1 of its parent.

For this study we propose a modified SDC metric which is calculated as follows;

a. The total number of unique elements
b. An extra point for every element that is mandatory
c. An extra point for every nesting level

- **Hybrid or "*goodness of fit*" Metrics**

  The following metrics represent the "goodness of fit" measure for data quality or determine whether a mapping is good or not.

  ii. DRR/SDC: A good mapping being determined by a low denominator and a high numerator. A low SDC represents less complexity from a project implementation perspective. This is always desirable. A high DRR implies our source data is more complete. Hence this hybrid ratio should reflect these facts as a way of assessing a good mapping.

  iii. MAR/SDC: A good mapping similarly determined as above.

.

CHAPTER 4
DESIGN AND IMPLEMENTATION

4.1 Source Data

The source data is derived from COMESA records and documents. In line with the theme of interoperability, the selection of this data was clearly biased towards the data that COMESA uses in its interaction with outside clients. For this reason, the data clusters selected were:

- Finance data – purchase order
- Human Resource Data – person details, postal address, employment history, education history
- Statistics – international trade statistics

An issue pertaining to this data when compared to that of counterpart standards is the fact that COMESA documents, such as the application form for the human resource department, contain several "types" of data that are presented in separate schemas in the comparable standards. For example, education history, employment history and postal address appear in one document but are presented as three different schemas in the HR standards we used for comparison and which are defined in chapter 4.

The purchase order, person name, postal address, employment history, education history and international trade statistics have 15, 3, 21, 13 and 14 elements respectively.

The data file used in the computation of the metrics is supplied with the tool discussed in section 4.4.

4.2 Destination Data

The choice of destination data was based on existing source data. A purchase order schema from the BASDA consortium of business standards was used to compare to the first data cluster in bullet point one above (BASDA, 2004)

The HR-XML consortium has an extensive selection of schema on HR based data. For comparison with COMESA data, we used the postal address, employment history, person name and education history as counterpart data to existing COMESA data (HR-XML Consortium, 2007).

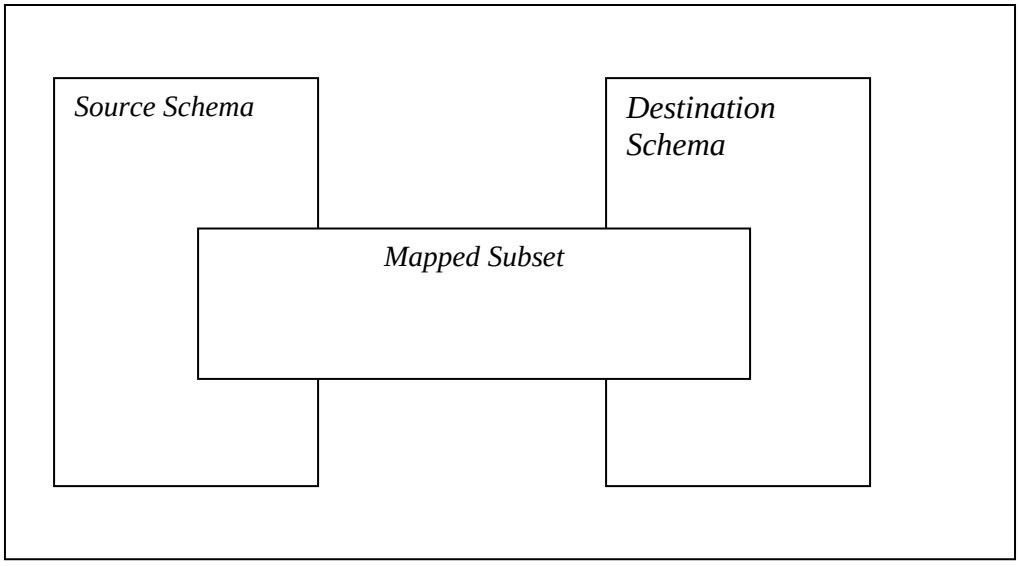
The SDMX schema on merchandise trade statistics was used for the statistical data comparison (SDMX, 2005).

Appendix 1 gives the details of the data fields in these schemas.

4.3 Data Mappings

The mapping is a process of linking source and destination elements that represent the same data. Schematically this is:

Figure 1: Data Mapping



Some preprocessing of data schemata was necessary prior to producing metrics. As the MS EXCEL spreadsheet program used does not explicitly indicate nesting levels of a schema for instance, some explicit indication in the spreadsheet was necessary in order to represent this. Further data cleaning

was necessary in order to ensure that fields with prefixes which represented elements were retained. Repeating fields were critical and retained but clearly indicated in the spreadsheet as such. In this regard the schema specification assisted in identification of these fields.

An example of this mapping is illustrated below for the postal address cluster from the HR-XML data standard. The middle column links COMESA data elements with the elements they map to in the data standard, i.e., column 3 data elements. This process required importation of data standard schemas and some preprocessing i.e. removal of unnecessary data from the imported data.

The final mapping file is shown in Table 6 below.

Table 6: Postal Data Mapping

COMESA Data	***Mapping***	**HR-XML Data**
Country	→	Country Code
Post Code	→	PostalCode
City	→	Municipality
Street Name	→	StreetName
PO Box	→	PostOfficeBox
		Region
		AddressLine
		BuildingNumber
		Unit
		Recipient
		AdditionalText
		Organization
		OrganizationName

4.4 The Tool for generation of metrics

In order to ease the computation of the several metrics considered in study a JAVA based tool was developed. This tool shields the user from the need to constantly use and adjust formulae in the spreadsheet once new data are evaluated. Using this tool, an evaluation of the COMESA data against the data schemas that are applicable to it was undertaken. Bearing in mind that the core objective of the research was the generation of metrics, the tool was really for the purpose of their computation. The data is imported from the spreadsheet into an MS Access table and a JAVA program used to compute the metrics. To connect JAVA to MS Access, we use an application program interface (API) called JAVA Database Connectivity (JDBC). This API enables the encoding of statements in Structured Query Language (SQL) that are passed to the program that manages the database. Since JDBC is similar to the SQL Access Group's Open Database Connectivity (ODBC), with a bridge program, we are able to use the JDBC interface to access databases through the ODBC interface. In addition to connectivity to the MS Access database fields, the program also inserts the computed metrics into another MS Access table for all data clusters. This makes the production of metrics a one-off exercise wherein all metrics for all clusters are computed. The tool's ability to interact with the MS Access table and then insert the computed metrics into another table allows the user easy access to the full set of metrics. Figures 2-5 illustrate some snapshots of the source and destination data as well as the outputs from the JAVA tool developed for computation of the metrics.

Figure 2: Source and Destination Data MS ACCESS Table

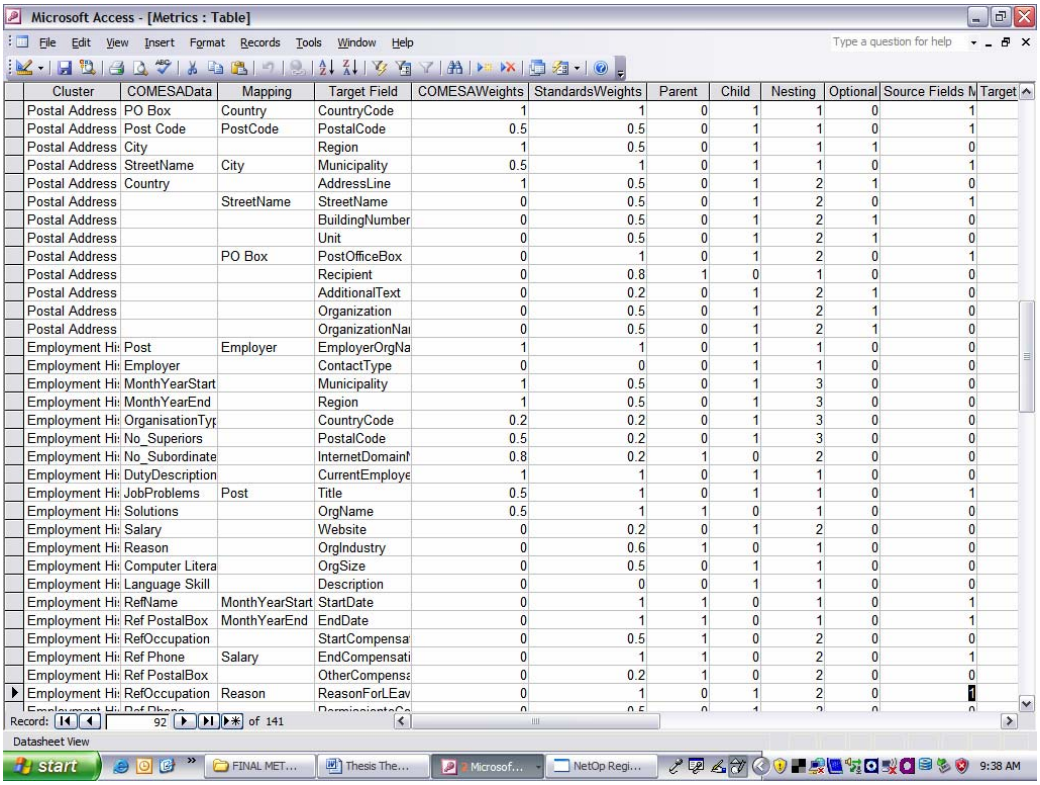


Figure 3: Java Tool Metrics Output 1

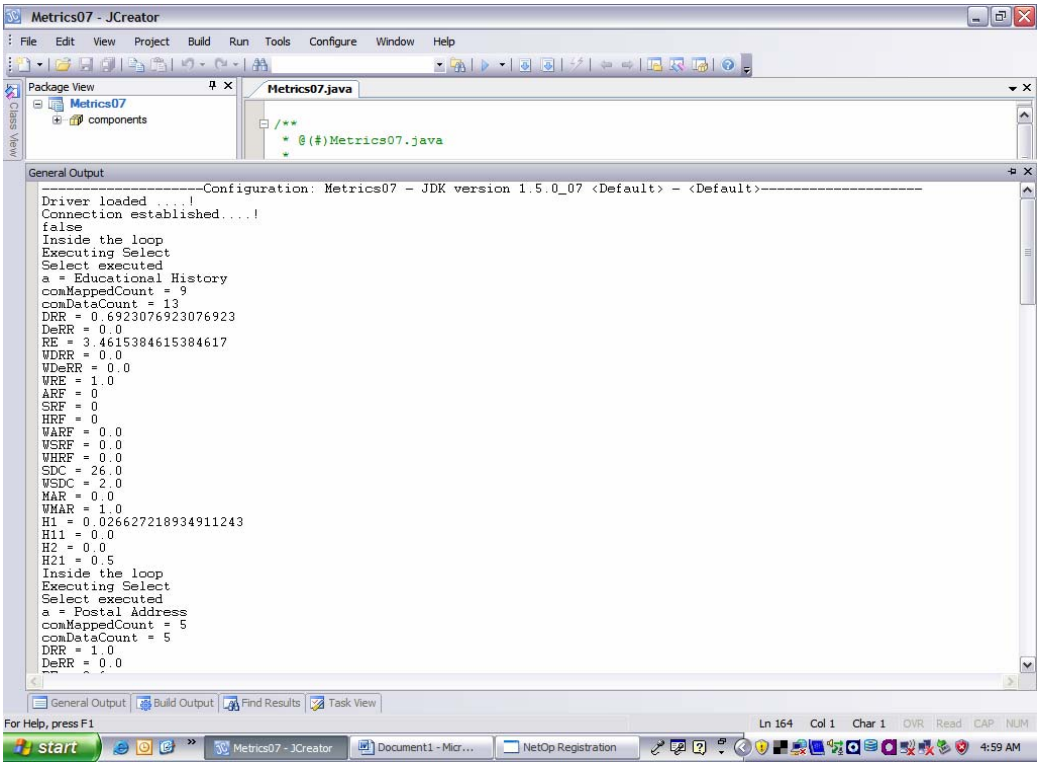


Figure 4: Java Tool Metrics Output 2

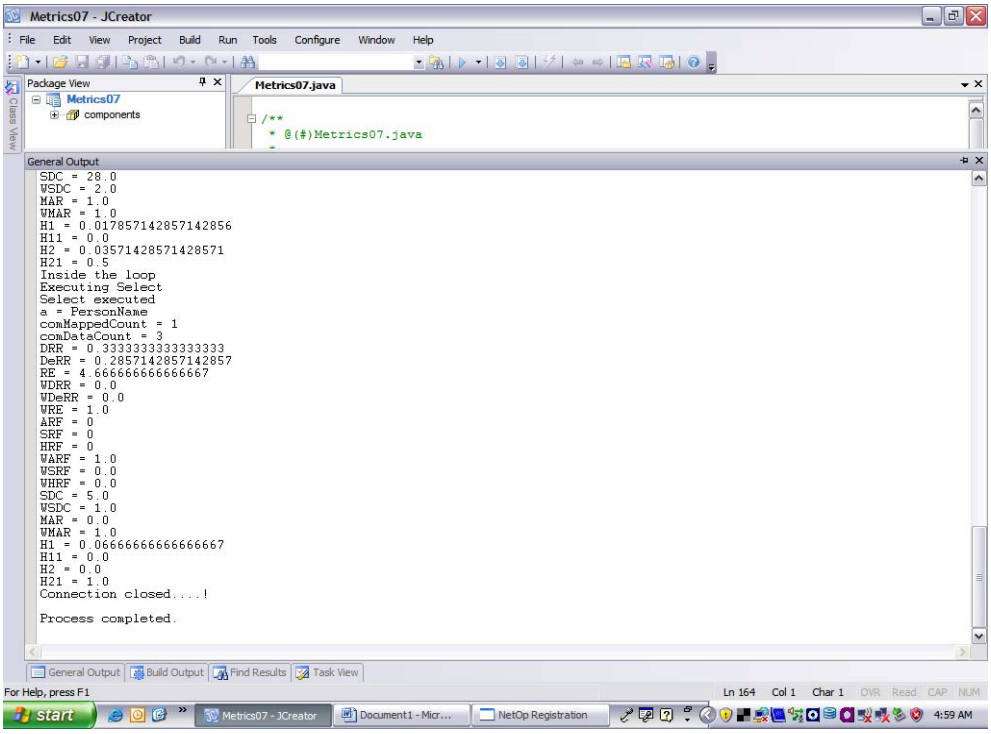
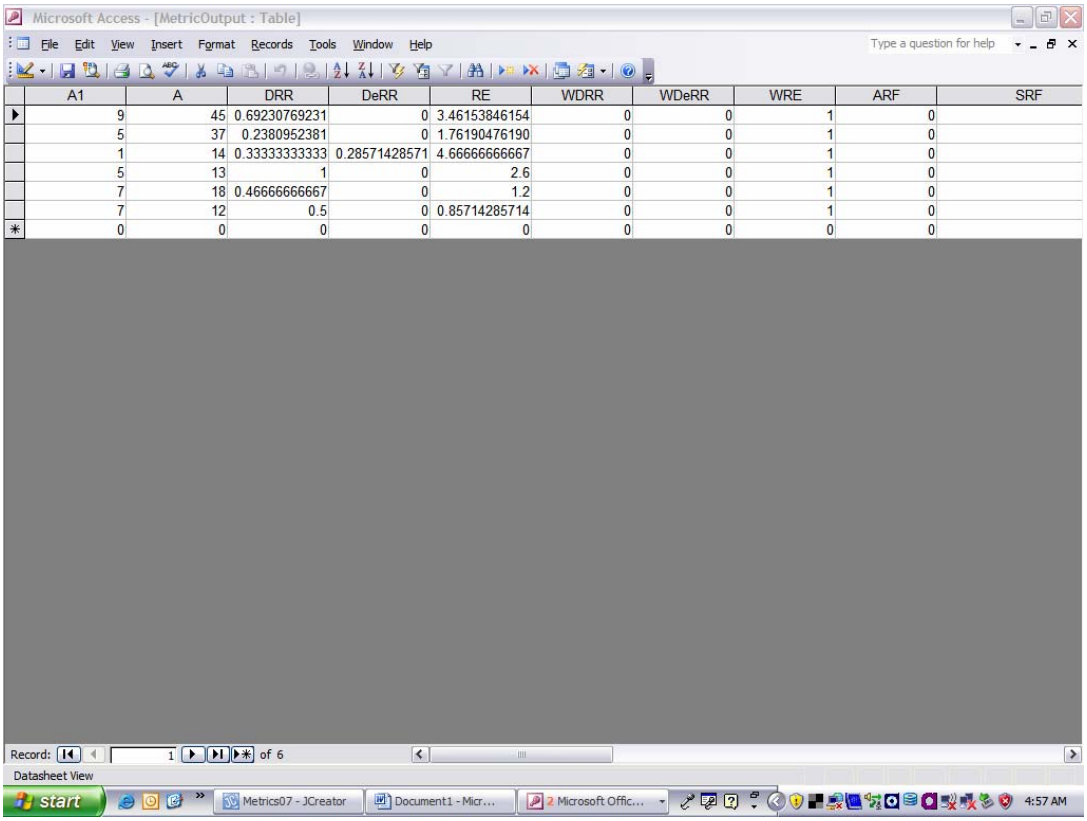


Figure 5: Metrics Output in MS ACCESS Table



A1	A	DRR	DeRR	RE	WDRR	WDeRR	WRE	ARF	SRF
9	45	0.69230769231	0	3.46153846154	0	0	1	0	
5	37	0.2380952381	0	1.76190476190	0	0	1	0	
1	14	0.33333333333	0.28571428571	4.66666666667	0	0	1	0	
5	13	1	0	2.6	0	0	1	0	
7	18	0.46666666667	0	1.2	0	0	1	0	
7	12	0.5	0	0.85714285714	0	0	1	0	
0	0	0	0	0	0	0	0	0	

CHAPTER 5
EVALUATION AND ANALYSIS

In the prior chapters, we discussed the research problem related to the assessment of COMESA organizational data using existing data standards. Chapter 3 laid out the methodology to be employed for this assessment and presented a number of quality metrics. In this chapter, an analysis of how the COMESA data conforms to international data standards is undertaken. Metrics are generated for the COMESA data to determine its conformance to these standards. Additionally, a second set of analyses incorporates expert user group's weighting factors as discussed in the methodological chapter.

In each section, an alternate set of similar metrics are presented incorporating expert users' input. This input is achieved through the introduction of weights to the fields/elements used in their computation. While the aforementioned metrics are straightforward to compute, they lack a reflection of how human beings measure the attributes the metrics represent. Both the source and destination are weighted by users. Users were representative of the type of data considered in the experiment. These were drawn from the finance, human resource and statistics departments. These departments were represented by five, three and two staff members respectively.

A weighting profile ranging from 0-1 was proposed to the expert users. The process and general objective of the experiment was explained to the users. Both the source and target data were explained as well as the weighting scheme. Each user then provided a weight for each element. The average of these weights for each element is presented in Appendix 1. This average was used for the computation of the weighted metrics. It is important to note that these weights are subjective. Hence the interpretation of results needs to take this into consideration.

5.1 Completeness Metrics

5.1.1 Unweighted Completeness Metrics

Six data clusters are used in this analysis. These data clusters are: Educational History, Employment History, Person Name, Postal Address, Purchase order and Trade Statistics. The completeness metrics are presented in Table 7 below.

In terms of relative numbers of elements, data elements in the data standards data number more than those represented in the COMESA data for all data clusters except the *TradeStatistics* cluster.

The data retention rate (DRR) is used to measure completeness. The *PostalAddress* data cluster has the highest DRR of 1, implying all the COMESA postal data are mapped. *EducationalHistory* and *PersonName* data clusters recorded DRRs of 0.69 and 0.67 respectively. The lowest DRR (0.32) was of the *EmploymentHistory* data.

Redundancy in the data is measured using the destination redundancy rate (DeRR). Only the *PersonName* data had a positive DeRR. It is the only data schema that had elements from the destination format that mapped to the same data elements in the source format.

Table 7: Unweighted Completeness Metrics

	Educational History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
Data Retention Rate	0.69	0.24	0.33	1.00	0.47	0.50
Destination Redundancy Rate	0.00	0.00	4.00	0.00	0.00	0.00
Relative Elements Metric	3.46	1.76	4.67	2.60	1.20	0.86

5.1.2 Completeness Metrics, Expert Users' Perspective

Table 8 below summarizes the weighted completeness metrics. Compared to the previous section on completeness metrics, the data retention rates for the *PersonName* and *Trade Statistics* data clusters increase as a result of applying expert user weights to cluster elements. This clearly implies more weight allocation to elements that are actually mapped and less weight to elements that are not.

The *Postal Address* cluster which recorded a DRR of 1 in Section 5.1.1 above now has a reduced DRR of 0.5. The incorporation of user weights clearly means that less weight has been given to mapped elements while more weight was given to those source data elements that are unmapped.

For relative element counts, the effect of the weighting of elements is a decrease for *EducationalHistory*, *PersonName*, *PostalAddress* and *PurchaseOrder* clusters. For these clusters, less weight was given by the expert users to elements in the destination schema that are not mapped. Hence the relative elements metric (REM) ratio reduces. *EmploymentHistory* and *TradeStatistics* clusters show increases in their REM. This reflects high weights assigned by users for elements in the destination schema which were unmapped.

Table 8: Weighted Completeness Metrics

	Educational History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
Data Retention Rate	0.57	0.08	0.67	0.50	0.40	0.66
Destination Redundancy Rate	0.00	0.00	4.00	0.00	0.00	0.00
Relative Elements Metric	2.12	2.25	4.20	2.00	1.10	1.18

5.2 Relevance Metrics

5.2.1 Unweighted Relevance Metrics

Relevance metrics are presented in Table 9 below. What is evident is that the *PersonName* data is the only cluster with an alias redundancy factor greater than 0. All other data clusters have no redundancies as they had no aliases, synonyms and homonyms.

Table 9: Unweighted Relevance Metrics

	Educational History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
Alias Redundancy factor	0	0	0.071	0	0	0
Synonym Redundancy factor	0	0	0	0	0	0
Homonym Redundancy factor	0	0	0	0	0	0

5.2.2 Relevance Metrics, Expert Users' Perspective

When weighted elements are considered, the Alias redundancy factor for the *PersonName* cluster increases to 0.159. The elements in the destination schema which are aliases have a high weight according to the expert users' perception. Table 10 below summarizes the metrics.

Table 10: Weighted Relevance Metrics

	Educational History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
Alias Redundancy factor	0	0	0.159	0	0	0
Synonym Redundancy factor	0	0	0	0	0	0
Homonym Redundancy factor	0	0	0	0	0	0

5.3. Concise Representation

Concise representation is measured using the ratio of information to markup. Table 11 below summarizes these metrics. A comparison of these ratios for COMESA data and data standards yields the following conclusion: the *EducationHistory*, *EmploymentHistory*, *PostalAddress*, *PurchaseOrder* and *TradeStatistics* data clusters have a more concise representation for COMESA data than their respective data standards. Only for the *PersonName* cluster is the HR-XML standard more concise that COMESA data.

Table 11: Concise Representation Metrics

		Education History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
COMESA Data	Total Information	70	48	49	28	45	34
	Total Markup	59	51	68	36	45	89
Data Standard	Total Information	56	45	49	24	45	34
	Total Markup	75	68	60	57	70	126
COMESA Data	Information/ Markup	1.19	0.94	0.72	0.78	1	0.38
Data Standard	Information/Markup	0.75	0.66	0.82	0.42	0.64	0.27

5.4 Complexity Metrics

5.4.1 Unweighted Complexity Metrics

The mappings additions ratio (MAR) is a metric that reflects the consequence of the addition of extra elements that would need to be supported by the COMESA IT staff. Typically the MAR is used to determine whether the adoption of a standard unchanged will have the consequence of adding extra elements that would need to be supported. Table 12 below summarizes these complexity metrics.
EducatonHistory, *EmploymentHistory*, *PersonName* and *PostalAddress* clusters are highly diverged from the ideal MAR ratio of 1:1. This implies the existence of extra capability that would need to be supported by COMESA if these standards are adopted unchanged. Do these correspond to the COMESA business requirements? We assess this through expert users' weighting of these metrics in Section 5.4.2 below.

Table 12: Unweighted Complexity Metrics

	Educational History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
Mappings Additions Ratio	**0.29**	**0.57**	**0.21**	**0.38**	**0.83**	**1.17**
SDC	**86**	**77**	**23**	**21**	**34**	**26**

The structured document complexity (SDC) metric is computed in order to give an indication of how complex a schema is from the perspective of a project's implementation. *EducationHistory* and *EmploymentHistory* clusters have higher SDC metrics implying high complexity in their possible implementation. Proceeding to further dissagregate the sources of complexity in the derivation of the SDC reveals clearly the number of elements as the most significant contributor to complexity.
The SDC metrics of *PersonName* and *PostalAddress* imply that, among the data clusters listed, they are the least difficult to implement. *PurchaseOrder* and *TradeStatistics* clusters are comparatively more difficult to implement. In all, the SDC metric is heavily dependent on the number of elements.

5.4.2 Complexity Metrics, Expert Users' Perspective

The effect of incorporating user input into the weighting of the metrics is a general increase in the MAR metric for all clusters except for the *PersonName* cluster which decreases slightly. Clearly the input of expert users implies that elements that are mapped receive higher weights than those that are unmapped hence a comparatively improved MAR metric. Table 13 below highlights these weighted complexity metrics.

The effect of the weighting on the SDC metric is significant reductions for all clusters. Particular mention needs to be made of the *EducationHistory* and *EmploymentHistory* whose SDC is reduced by over 49 percent and 59 percent respectively.

These reductions are due to the fact that these metrics are heavily dependent on the number of elements, which when expert users' weighting is considered, skew more towards some elements than others.

Table 13: Weighted Complexity Metrics

	Educational History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
Mappings Additions Ratio	0.47	0.45	0.24	0.50	0.91	0.85
SDC	44	31	15	18	31	22

5.5 Hybrid metrics representing the "goodness of fit" of a mapping

5.5.1 Unweighted Hybrid Metric

The following hybrid metrics are presented in order to answer, describe and quantify how good a mapping is. In order to derive these hybrid metrics, we combine metrics for which a higher numerator and lower denominator is desirable. This leads to a metric for which higher values represent a good mapping or better mapping comparatively.

The hybrid metrics we use are the ratio of the DRR to the SDC and the ratio of the MAR to the SDC and these are summarized in Table 14 below.

With regard to the MAR/SDC metric, and ordering the clusters by better mapping, we have the following order: *TradeStatistics*, *PurchaseOrder*, *PostalAddress, PersonName* , *EmploymentHistory* and *EducationHistory*. For the DRR/SDC, the order is as follows: *PostalAddress, TradeStatistics*, *PurchaseOrder*, *PersonName* , *EducationHistory* and *EmploymentHistory.*

Table 14: Unweighted Goodness of Fit Metrics

	Educational History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
DRR/SDC	0.008	0.003	0.014	0.048	0.014	0.019
MAR/SDC	0.003	0.007	0.009	0.018	0.025	0.045

5.5.2 Goodness of Fit Metrics, Expert Users' Perspective

Table 15 below provides an overview of weighted goodness of fit metrics. In terms of ranking of clusters, the DRR/SDC ratio places the *PersonName* cluster as the best fit of all clusters. This is followed by the *TradeStatistics* and *PostalAddress* clusters. The MAR/SDC ratio has the *TradeStatistics*, *PurchaseOrder* and *PostalAddress* as clusters with the best mapping.

How does the weighting affect the magnitude of the ratios such that improvements/reductions are observed among these data clusters? The weighting yields better, i.e. higher hybrid ratios in the case of the DRR/SDC metric for *EducationalHistory*, *PersonName*, *TradeStatistics* while for the MAR/SDC metric, higher metrics are observed for all clusters except the *TradeStatistics* cluster.

Table 15: Weighted "Goodness of Fit" Metrics

	Educational History	Employment History	Person Name	Postal Address	Purchase Order	Trade Statistics
DRR/SDC	0.013	0.002	0.046	0.028	0.013	0.030
MAR/SDC	0.011	0.014	0.016	0.028	0.030	0.039

5.6. Analysis

In this chapter quality metrics were generated that facilitated comparison of COMESA data with existing data standards from the XML-HR Consortium, the BASDA data standards and the SDMX statistical data standards.

The completeness attribute of the data represents the extent to which data is not missing and is of sufficient breadth and depth for the task at hand. Our DRR metric suggested that for the *PostalAddress, EducationHistory and PersonName* data clusters COMESA data were mapped at 100%, 69% and 50% respectively. The incorporation of expert users' weights changed this order to yield a scenario where the most mapped data clusters were *PersonName, TradeStatistics* and *EducationHistory* with retention rates of 67%, 66% and 57% respectively.

The relative elements ratios suggested that the data standards used above had more elements than are represented in COMESA data.

For concise representation, a comparison of the information to markup ratios for COMESA data and data standards found that almost all data clusters had a more concise representation for COMESA data than their respective data standards. The exception was the *PersonName* data cluster which was more concise for the HR-XML standard than COMESA data.

The analysis also looked at the generation of complexity metrics. For these the MAR and SDC metrics were generated. For the MAR, it was observed from the data that *EducatonHistory*, *EmploymentHistory*, *PersonName* and *PostalAddress* clusters displayed MARs that are diverged from 1:1, clearly implying the existence of extra capability that would need to be supported by COMESA if these standards were adopted unchanged. The incorporation of user input into the weighting of the metrics led to a general increase in the MAR metric for all clusters except for the *PersonName* cluster which decreased slightly. This implied that elements that were mapped received higher weights than those that were unmapped hence a comparatively improved MAR metric.

The SDC metric suggested greatest difficulty in the implementation of the *EducationHistory* and *EmploymentHistory* data clusters than other clusters, a scenario that remained unchanged when expert users' input was incorporated in the computation of the metric.

Which mapping was the best? In order to determine this, "goodness of fit" metrics were generated. These metrics were ratios of the DRR to SDC and the MAR to SDC. These metrics both suggested

better mappings of COMESA data to related standards for the *TradeStatistics*, *PurchaseOrder* and *PostalAddress* data clusters than for *EducationHistory*, *EmploymentHistory* and *PersonName* data clusters. What determined this result? For both metrics, we attribute this to the complexity or specifically number of elements in the data standards which were used relative to the COMESA data elements.

When we incorporated expert users' inputs in the metric computations, the DRR/SDC metric suggested the best mapping for the *PersonName* data cluster, while *TradeStatistics* and *PostalAddress* were second and third. For the MAR/SDC metric, *TradeStatistics*, *PurchaseOrder* and *PostalAddress* data clusters had the best mappings.

5.7 Implementation Issues and Recommendations

There are a number of implications arising from the experiment conducted in this study. First, does the source data rely on data sources for which the international standards are being promoted? The interaction between various organizations that necessitates the use of standards is what determines the answer to this question. With regard to the source data that was analyzed in this study, two categories of data sources emerge, local and international. The standards discussed are international and as such the lack of a "national" contribution to this standards process has an impact on the degree to which source data is mapped onto the target standards data. COMESA would have to undertake an inventory of how its information systems discussed in this paper rely on data sources for which standards are being promoted and also how much it relies on data sources for which standards are not being promoted.

Secondly the study identified users of this data within the organizational context. However there is a need to identify both users and collectors of the data in order to assess who would be the beneficiary of a standards-based information system.

Thirdly and from an organizational perspective there is need to designate a team responsible for creation of this standards-based information system. Such a team would be vested with the responsibility of:

- Identification of priority questions to be answered by standards-based system.
- Ensuring management buy in.

- Selection of relevant data elements needed to answer bullet point 1 above.
- Decision on how data is to be organized.
- Determination of necessary response time required for the system.
- Benchmark the priority questions against financial implications in order to determine the scope of resulting system design and development costs.

Fourthly, the adoption of any standards implies the need for involvement in consortia that are responsible for them. This implies dedication of resources towards initiation of a discussion regarding a business case for incorporating the specific data requirements into the standard. This is particularly pertinent given the completeness metrics produced in this study.

CHAPTER 6

CONCLUSION

6.1 Conclusion

The study's objective was to demonstrate the extent to which COMESA's data maps onto international data standards through the generation of metrics. The technique adopted in this study is feasible for two reasons. First the technique adopted in this research entailed a review of both source and target data in order to select appropriate formats to enable their comparison.
Secondly the definition of metrics was such that it allowed easy response to immediate critical questions in an environment where feasibility of a move to standardization is sought. Some of these critical questions were as follows:

- How much of my source data is mapped?
    - Or how complete is a mapping of a COMESA document compared to the data standard?
- How complex is the standard's schema?
- Are there extra elements the COMESA documents would have to support?

The ensuing metrics were an effective way of getting answers to these questions and provided a basis for further organizational decision for standardization as the previous chapter demonstrates. Hence a standardization transition is clearly informed by the observations which relate directly to the findings on the attributes the metrics measured in the study. The need and significance of interoperability notwithstanding, the study successfully demonstrated the use of specific data quality metrics to demonstrate the feasibility of achieving it. Further, with regards to achieving interoperability and the bullet point three above, the study also demonstrated scope to assess whether full or partial interoperability is the practical and desired goal of the organization.

6.2 Future Work

This study essentially considered the feasibility of a move to standardization for a single organization. More research is needed in order to provide a fuller picture of the implications. Hence future work would have to consider:

- As quality is not an absolute value and surrogate measures are used to measure it, the results in this study may vary with different user communities so there is need for a concordance or translation between results generated by one set of users and another. This would essentially cater for the issue raised above by taking into account users and producers of the data. A concordance would be most beneficial in understanding the implications of a standardization transition.
- Linked to this, the results generated in this study would also have to be assessed by a human assessor in order to determine the extent of their usefulness.
- Standardization transitions have a clear cost implication. Hence development of tools for analysis and costing of transitions is an important future research work.
- As our technique is focused on a single organization. However future work needs to make comparisons with other measures and techniques for assessing standardization.
- Similar to the approach of Ochoa and Duval (2005), there is need to assess the quality of standards data using metrics. This is particularly important given that use of standards could easily take for granted their quality. Once this is known, the evaluation of source data against a standards-based data needs to incorporate their inherent quality for better informed assessments.

References

Business Application Software Developers, 2004, 'eBIS-XML version 3.09', viewed 3 July 2007, <http://www.basda.org/VD65/default.asp?PSID=51>

Comsys Information Technology Services Inc., 2002, 'Comsys-TIM System Metrics', viewed 6 August 2007, <http://www.comsysprojects.com/SystemTransformation/tmmetricguide.htm>

Chemical Industry Data Exchange, 2004, 'Chem eStandards v4.0 Message Standards Complete Edition', viewed July 2007, <http://www.cidx.org/Default.aspx?tabid=78>.

Dublin Core Metadata Initiative, 2008, 'Dublin Core Metadata Element Set, Version 1.1', viewed 7 February 2008, http://www.dublincore.org/documents/dces/.

European Industry Association, 2004, *EICTA Interoperability White Paper*, Brussels.

Human Resource-XML Consortium, 2007, 'HR-XML 2.5', viewed 3 January 2008, <www.hr-xml.org>.

ISO/IEC Guide 2, 1996; Definition 3.2.

International Telecommunications Union, 2007, 'ITU-T Recommendations', viewed 3 January 2008,< http://www.itu.int/ITU-T/publications/recs.html >.

Institute of Electrical and Electronics Engineers, 1990, IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. New York, NY.

Institute of Electrical and Electronics Engineers, 1998, IEEE Standard 1061-1998, Standard for a Software Quality Metrics Methodology, revision, Piscataway, NJ.

International Telecommunications Union, 1993, 'Information Technology Digital Compression and Coding of Continuous-Tone Still Images – Requirements and Guidelines, Recommendation T.81', viewed 7 February 2008, <http://www.w3.org/Graphics/JPEG/itu-t81.pdf>

Hughes T, 2006, 'The Challenge of Data Interoperability from an Operational Perspective', viewed 4 July 2007, <http://iswc2006.semanticweb.org/program/panel-todd-huges.ppt#256,1>

International Portal on Food Safety, Animal & Plant Health, 1994, Sanitary and Phytosanitary Measures Agreement, viewed 3 January 2008, <http://www.ipfsaph.org/En/default.jsp>.

Jelliffe, R 2006, 'Metrics for XML Projects #3: XML Mapping Completeness Ratio', viewed 6 August 2007, <http://www.oreillynet.com/xml/blog/2006/05/metrics_for_xml_projects_3_xml_1.html>.

Jelliffe, R 2006, 'Metrics for XML Projects #1: Element and Attribute Count', viewed 6 August 2006, <http://www.oreillynet.com/xml/blog/2006/05/metrics_for_xml_projects_1_ele.html>

Jelliffe, R 2006, 'Metrics for XML Projects #4: XML Mapping Additions Ratio', viewed 6 August 2006, <http://www.oreillynet.com/xml/blog/2006/05/metrics_for_xml_projects_4_xml_1.html>.

Jelliffe, R 2006, 'Metrics for XML Projects #5: Structured Document Complexity Metric' viewed 6 August 2006,
<http://www.oreillynet.com/xml/blog/2006/05/metrics_for_xml_projects_5_str_1.html>

Kaner, C & Bond W, 2004, 'Software Engineering Metrics: What Do They Measure and how do we know?', paper presented to the 10th International Software Metrics Symposium, Chicago, 2004.

Loshin, D 2005, 'Monitoring Data Quality Performance using Data Quality Metrics', viewed 6 August 2007,
<http://www.it.ojp.gov/documents/Informatica_Whitepaper_Monitoring_DQ_Using_Metrics.pdf >

OASIS, 2006, 'DocBook v4.5', viewed 3 July 2007, <http://www.oasis-open.org/specs/index.php#dbv4.5>

OASIS, 2002, 'Collaboration-Protocol Profile and Agreement Specification Version 2.0', viewed 7 February 2008, <http://www.oasis-open.org/committees/ebxml-cppa/documents/ebcpp-2.0.pdf>.

Ochoa, X & Duval, E, 2005, 'Towards Automatic Evaluation of Metadata Quality in Digital Repositories', viewed 27 July 2007,
<http://ariadne.cti.espol.edu.ec/M4M/files/TowardsAutomaticQuality.pdf>

Okawara T, Tanaka J, Morishima A & Sugimoto S, 2005, 'A Support Tool for XML Schema Matching and Its Implementation', paper presented at the Proceedings of the 21st International Conference on Data Engineering, Tokyo, 2005.

Open Directory Project, 2007, 'Protocols', viewed 10 July 2007,
<http://www.dmoz.org/Computers/Internet/Protocols/>

Open Source Initiative, 2006, 'The Open Source Definition', viewed 10 July 2007,
<http://www.opensource.org/docs/definition.php>

Open Directory Project, 2007, 'Data Formats', viewed 10 July 2007,
<http://www.dmoz.org/Computers/Data_Formats/>

Open Travel Alliance, 2007., 'OpenTravel 2007B" viewed 3 July 2007,
<http://www.opentravel.org/Specifications/Default.aspx>

Pipino L Lee Y & Wang R, 2002, 'Data Quality Assessment', *Communications of the ACM,* Vol. 4.

Roelofs G 2007, 'Portable Network Graphics', viewed 3 July 2007, <http://www.libpng.org/pub/png/>

Statistical Data and Metadata Exchange, 2005, 'SDMX-ML: Schema and Documentation (Version 2)', viewed 3 July 2007, <http://www.sdmx.org/docs/2_0/SDMX_2_0_SECTION_03A_SDMX_ML.pdf>

Stephens T, 2005, 'Knowledge: The Essence of Meta Data: Metrics and the Source of all Knowledge', *DM Review Online,* viewed 6 August 2007, <http://www.dmreview.com/news/1037094-1.html>

World Health Organization, 2005, Food standards (Codex Alimentarius), viewed 3 January 2008, <http://www.who.int/foodsafety/codex/en>

World Wide Web Consortium, 1999, 'HyperText Markup Language', viewed 5 January 2008, <http://www.w3.org/TR/html4>

World Wide Web Consortium, 2004, 'Resource Description Framework', viewed 7 February 2008, <http://www.w3.org/TR/REC-rdf-syntax/>

World Wide Web Consortium, 2006, 'Extendible Markup Language', viewed 7 February 2008, <http://www.w3.org/TR/2006/REC-xml-20060816/>

Appendix 1

Cluster	COMESA Data	Average COMESA Weights	Destination Data Fields	Average Destination Data Weights
Educational History	University Name	1.00	InternetDomainName	0.50
Educational History	City	0.80	School Id	0.40
Educational History	Country	**1.00**	School Name	**1.00**
Educational History	StartDate	1.00	Municipality	0.80
Educational History	EndDate	1.00	Region	0.50
Educational History	Degree	1.00	CountryCode	1.00
Educational History	MainCourse	0.80	PostalCode	0.20
Educational History	School Name	0.50	OrganizationUnit	0.40
Educational History	City	0.40	Attendance Status	1.00
Educational History	Country	0.40	Degree Type	1.00
Educational History	StartDate	0.60	Exampassed	1.00
Educational History	EndDate	0.60	Graduating Degree	1.00
Educational History	Certificate	0.80	Degree Name	1.00
Educational History		0	Academic honours	0.50
Educational History		0	honorsProgram	0.50
Educational History		0	DegreeDate	0.80
Educational History		0	OtherHonors	0.20
Educational History		0	DegreeMajor	0.50
Educational History		0	Program ID	0.20
Educational History		0	DegreeConcentration	0.50
Educational History		0	Name of major	0.80
Educational History		0	Option	0.50
Educational History		0	DegreeMinor	0.50
Educational History		0	Program ID	0.20
Educational History		0	Name	0.20
Educational History		0	Degree Measure Type	0.20
Educational History		0	*MeasureSystem*	0.20
Educational History		0	*MeasureValue*	0.20
Educational History		0	*LowestPossibleValue*	0.20
Educational History		0	*HighestPossibleValue*	0.20
Educational History		0	*ExcessiveValueIndicator*	0.20
Educational History		0	*GoodStudentIndicator*	0.20
Educational History		0	AcademicCredit Code	0.40
Educational History		0	CourseLevelCode	0.20
Educational History		0	CumulativeSummaryIndicator	0.20
Educational History		0	Academic CreditHoursIncluded	0.20
Educational History		0	AcademicCreditHoursAttempted	0.20
Educational History		0	AcademicCreditHoursEarned	0.20
Educational History		0	numberofStudents	0.00
Educational History		0	Enrolment Status	0.20
Educational History		0	CurrentlyEnrolled	0.20
Educational History		0	StudentInGoodStanding	0.50
Educational History		0	StartDate	1.00
Educational History		0	EndDate	1.00

Educational History		0	Comments	0.10
PersonName	Family Name	1.00	FormattedName	1.00
PersonName	First Name	0.50	LegalName	1.00
PersonName	Maiden Name	0	GivenName	1.00
PersonName		0	PreferredGivenName	0.10
PersonName		0	MiddleName	0.50
PersonName		0	FamilyName	1.00
PersonName		0	Affix	1.00
PersonName		0	FormattedName3	0.10
PersonName		0	LegalName4	0.10
PersonName		0	GivenName5	0.10
PersonName		0	PreferredGivenName6	0.10
PersonName		0	MiddleName7	0.10
PersonName		0	FamilyName8	0.10
PersonName		0	Affix11	0.10
Postal Address	PO Box	1.00	CountryCode	1.00
Postal Address	Post Code	0.50	PostalCode	0.50
Postal Address	City	1.00	Region	0.50
Postal Address	StreetName	0.50	Municipality	1.00
Postal Address	Country	1.00	AddressLine	0.50
Postal Address		0	StreetName	0.50
Postal Address		0	BuildingNumber	0.50
Postal Address		0	Unit	0.50
Postal Address		0	PostOfficeBox	1.00
Postal Address		0	Recipient	0.80
Postal Address		0	AdditionalText	0.20
Postal Address		0	Organization	0.50
Postal Address		0	OrganizationName	0.50
Employment History	Post	1.00	EmployerOrgName	1.00
Employment History	Employer	0.00	ContactType	0.00
Employment History	MonthYearStart	1.00	Municipality	0.50
Employment History	MonthYearEnd	1.00	Region	0.50
Employment History	OrganisationType	0.20	CountryCode	0.20
Employment History	No_Superiors	0.50	PostalCode	0.20
Employment History	No_Subordinates	0.80	InternetDomainName	0.20
Employment History	DutyDescription	1.00	CurrentEmployer	1.00
Employment History	Job Problems	0.50	Title	1.00
Employment History	Solutions	0.50	OrgName	1.00
Employment History	*Salary*	0	Website	0.20
Employment History	*Reason*	0	OrgIndustry	0.60
Employment History	Computer Literacy	0	OrgSize	0.50
Employment History	Language Skill	0.00	Description	0.00
Employment History	RefName	0	StartDate	1.00
Employment History	Ref PostalBox	0	EndDate	1.00
Employment History	RefOccupation	0	StartCompensation	0.50
Employment History	Ref Phone	0	EndCompensation	1.00

Employment History	Ref PostalBox	0	OtherCompensation	0.20
Employment History	RefOccupation	0	ReasonForLEaving	1.00
Employment History	Ref Phone	0	PermissiontoContact	0.50
Employment History		0	VerifiyEmployment	0.50
Employment History		0	EligibleforRehire	0.20
Employment History		0	*AttendanceRating*	0.20
Employment History		0.00	NumericValue	0.00
Employment History		0.00	StringValue	0.00
Employment History		0.00	OverallPerformanceRating	0.80
Employment History		0.00	NumericValue	0.00
Employment History		0.00	StringValue	0.00
Employment History		0	QuestionAnswerPair	0.10
Employment History		0	Question	0.10
Employment History		0	Answer	0.10
Employment History		0	JobPlan	0.10
Employment History		0	JobGrade	0.10
Employment History		0	JobStep	0.10
Employment History		0	Comments	0.10
Employment History		0	JobCategory	0.10
PurchaseOrder	Date	1.00	order	0.93
PurchaseOrder	name	1.00	Orderhead	0.65
PurchaseOrder	postalcode	0.40	OrderReferences	0.80
PurchaseOrder	street	0.70	Extensions	0.40
PurchaseOrder	city	0.80	OrderDate	1.00
PurchaseOrder	OrderNo	1.00	Supplier	1.00
PurchaseOrder	Item No.	0.83	Originator	0.70
PurchaseOrder	Quantity	1.00	Buyer	0.78
PurchaseOrder	Description	1.00	Delivery	0.73
PurchaseOrder	UnitPrice	1.00	InvoiceTo	0.83
PurchaseOrder	Amount	1.00	OrderLine	1.00
PurchaseOrder	Total	1.00	PercentDiscount	0.60
PurchaseOrder	Division	0.78	AmountDiscount	1.00
PurchaseOrder	OrderBy	0.70	SpecialInstructions	0.60
PurchaseOrder	ApproveBy	0.83	Narrative	1.00
PurchaseOrder		0	Settlement	0.65
PurchaseOrder		0	TaxSubTotal	0.70
PurchaseOrder		0	OrderTotal	1.00
Trade Statistics	Year	1.00	CL_Decimals	0.70
Trade Statistics	Transport	0.85	CL_FREQ	1.00
Trade Statistics	HS1996	1.00	CL_ISO_CURRENCY	0.85
Trade Statistics	HS2002	1.00	CL_UN_COUNTRY	1.00
Trade Statistics	Country	1.00	CL_UN_DATASET STATUS	0.50
Trade Statistics	Local Currency	0.55	CL_UN_OECD_COMMODITY_HS2002	1.00
Trade Statistics	Dollar Value	0.55	CL_UN_QUANTITY_UNIT	0.75
Trade Statistics	Flow	0.75	CL_UN_TRADE_FLOW	1.00
Trade Statistics	Reporter Country	0.75	CL_CLASSIFICATION	0.55
Trade Statistics	HS Heading	1.00	CL_UN_TRADE-SYSTEM	1.00

Trade Statistics	HS Chapter	0	CL_UN_VALUATION	0.85
Trade Statistics	Quantity	0	CL_UNIT_MULT	0.75
Trade Statistics	NetWeight			
Trade Statistics	TradeType			