POLITECNICO DI MILANO

Master of Science in Computer Systems Engineering

Dipartimento di Elettronica e Informazione

A MODEL FOR ASSESSING THE VALUE OF INFORMATION INTEGRATION

Author: **Marcio Fossa VICENTINI (749156)**

Supervisor: **Prof. Chiara FRANCALANCI**

Supervisor: **Ing. Cinzia CAPPIELLO**

Date: **15/11/2011**

Academic Year: **2010 – 2011**

ABSTRACT

Issues with combining heterogeneous data sources under a single query interface have existed for quite some time. The rapid adoption of databases after the 1960s naturally led to the need to share and merge existing repositories. Among the innumerous areas of study related to information technologies, data integration is the area that focuses on combining data residing in different sources and providing users with a unified view of these data. The purpose of this thesis is to discuss a model that aims to unify several concepts diffusely investigated in the information systems literature, namely information quality, structure, diffusion, capacity, utility, and value. Through this model we intend to assess the value of information integration initiatives. In order to provide the reader with a more pragmatic approach to this topic, we intend to apply the concepts studied to a real-world practical scenario. The case study here presented is based on experiences carried out in the Mobile Services for Agro-food project that aims to develop value added services for the agro-food sector. It's important to notice that the issues considered in this paper have practical relevance for business in strategic and operational dimensions.

Key words: Databases, data integration, information, Mobile Services for Agro-food, quality, capacity, cost, utility, value, model.

SOMMARIO

Problemi con la combinazione di fonti di dati eterogenee sotto un'interfaccia unica di esecuzione di query esistono da parecchio tempo. La rapida adozione di basi di dati dopo il 1960 naturalmente ha portato alla necessità di condividere e fondere diversi repositori esistenti. Tra le innumerevoli aree di studio legate alla tecnologia dell'informazione, l'integrazione dei dati è l'area che si concentra sulla combinazione di dati residenti in fonti diverse in modo ad offrire agli utenti una visione unificata dei dati. Lo scopo di questa tese è quello di discutere un modello che mira a unificare diversi concetti diffusamente studiati nella letteratura dei sistemi informativi, cioè la qualità, la struttura, la diffusione, la capacità, l'utilità e il valore dell'informazione. Attraverso questo modello vogliamo analizzare il valore di iniziative di integrazione dell'informazioni. Per offrire al lettore un approccio un può più pragmatico di questo tema, applicheremo i concetti studiati in un scenario pratico del mondo reale. Il caso di studio qui presentato si basa sul l'esperienze svolte presso il progetto Mobile Services for Agro-food che si propone di sviluppare servizi a valore aggiunto per il settore agro-alimentare. È importante notare che i temi considerati in questo documento hanno rilevanza pratica per le imprese in dimensioni strategiche e operative.

Parole chiavi: Basi di dati, integrazione di dati, informazioni, Mobile Services for Agro-food, qualità, capacità, costi, utilità, valore, modello.

ACKNOWLEGMENTS

To my parents, Marcio and Eliete, and my brother, Murilo, who always believed in my potential and always provided support for my accomplishments.

To my friends, too many to name here, but that in many ways supported me and shared with me two amazing years here in Italy.

To my supervisors, Prof. Francalanci and Ing. Cappiello, for all the teachings, patience, and time devoted to help me develop this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Ms4A	Mobile Services for Agro-food
DI	Data Integration
EII	Enterprise Information Integration
LAV	Local as View
GAV	Global as View
DBMS	Database Management Systems
ETL	Extract, Transform, and Load
OLTP	On-Line Transaction Processing
OLAP	On-Line Analytical Processing
FDBMS	Federated Database Management Systems
WFMS	Workflow Management Systems
P2P	Peer-to-peer
CRM	Customer Relationship Management
IIC	Intensional Information Capacity
EIC	Extensional Information Capacity
SOA	Service Oriented Architecture

1. INTRODUCTION

With a compound annual growth rate of almost 60%, the digital universe is growing faster and is projected to be nearly 1.8 zettabytes (1.8 trillion gigabytes) in 2011. In this scenario, the picture related to the source and governance of digital information is nothing but intriguing: According to researches, approximately 70% of the digital universe is created by individuals, yet enterprises are responsible for the security, privacy, reliability, and compliance of 85% of the information (Gantz, 2008).

Among the innumerous reasons that help explain this overwhelming growth of information, it's worth mentioning a few:

- **Internet:** Network of networks that consists of millions of private, public, academic, business, and government networks, of local to global scope, that are linked by a broad array of electronic, wireless and optical networking technologies.
- **Web 2.0:** Web applications that facilitate participatory information sharing, interoperability, user-centered design, and collaboration on the World Wide Web (O'Reilly, 2005).
- **Business intelligence data:** Organizations are increasingly interested in collecting social data and aggregating with third party information in order to support more targeted business decisions. In that sense, long-term retention of information is necessary in order to identify trends and predict future behavior.

Now, when we stop and think about the dimensions of this digital universe and all the possibilities that lie within, one cannot help but wonder: Are we making the best use of all the information that we have access to? How much more can we extract from this gigantic source of information? How do we tame chaos and extract the most value form it?

Among the many studies that have been conducted in the last few years, in order to answer these questions and provide reliable solutions to this issue, literature shows that the data integration theory has proved to be a very effective approach in the quest to extract more value from the information at our disposal.

In order to be useful to the users, information must be treated, as so to guarantee its quality, and also needs to be brought together (connected) in order to increase its relevance and accessibility. In a corporate environment, data integration initiatives boil down to money. While building standalone applications may be the "quick" solution to an immediate business problem, maintaining the resulting redundant and inconsistent databases and applications is a huge cost to organizations (Moss & Adelman, 2000).

According to (Moody & Walsh, 1999) there are seven laws of information: 1) information is infinitely sharable; 2) the value of information increases with use; 3) information is perishable; 4) the value of information increases with accuracy; 5) the value of information increases when combined with other information (integration); 6) more information is not necessarily better information (pertinence); 7) information is not depletable.

In this thesis, as well as gaining a better understanding on how data integration works, we also want to understand what are its impacts to information. In the following pages we will present innumerous concepts on data integration and information variables. In the end of this study, we expect to apply the knowledge acquired to a real-world case scenario in order to analyze in a more pragmatic way how to evaluate the value of information integration.

1.1. Thesis Overview

In order to gain a more global understanding of how this study will be structured, the following paragraphs intend to summarize the concepts and procedures that will be covered/adopted in the following chapters of this work. The remainder sections of this thesis are organized as follows:

Chapter 2 – Goals and Motivations: As to guide our efforts in the development of this thesis, this chapter presents the main motivations and goals of this study. Also, for future reference, we will also specify the study case scenario in which we will apply the concepts later viewed in this work.

Chapter 3 – State of the art: This chapter covers a variety of topics that are relevant to the project in question. Background information acquired through literature review in the field of data integration will be presented here. By the end of this chapter we hope the reader will have a better understanding of what is data integration, what are its issues, what are its benefits and finally, which are the most common approaches for data integration nowadays. The final section of this chapter presents a miscellaneous of topics that are necessary for building a basic vocabulary that will most probably be requested in following chapters of this thesis.

Chapter 4 - Methodology: This chapter deals with the execution basis of the project. It covers the methods that will be used to implement the project, more specifically, the first section of this chapter describes the roadmap of this project; the second and third sections present a unified quality based model for information value and capacity analysis; the fourth section presents a didactical example of schema integration methodology; and finally, in order to bridge everything exposed so far, a section on the connections between the theory developed and the Ms4A project.

Chapter 5 – Empirical Validation: This chapter walks the reader through the application of the model in the Ms4A context, as well as it analyzes all the relevant parameters proposed in the model, namely, value, capacity, quality and cost of information.

Chapter 6 – Discussion: This chapter proposes a discussion of the results presented on the previous chapter, not only by simply re-stating the reported results but by critically trying to extract what are the implications of the results obtained and what were the main issues in the project execution.

Chapter 7 – Conclusion: This chapter concludes the work done in this project, by synthesizing the most remarkable findings and by specifically answering to the project's established goals.

Chapter 8 – Recommendations: Finally, the last chapter proposes a look into the future and suggests directions for further researches as to fill in the gaps in our understanding of the still innumerous issues related to data integration.

2. MOBILE SERVICES FOR AGRO-FOOD: GOALS AND MOTIVATIONS

2.1. Mobile Services for Agrofood

In the last years there has been an increased interest in issues related to health and diet, both by consumers, and scientific community. The problems related to intolerance, allergies or diseases such as overweight and obesity are rapidly increasing. Consumers are increasingly careful about what they eat, and on the other hand food companies must offer products that are in line with new market demands in order to be able to satisfy their customers (OECD, 2004).

For these reasons, the consumers' interest towards food knowledge is growing, and in this context an important role is played by food labels. The role of labeling is to help consumers making the optimum choice during their purchasing activities and thus reducing information asymmetry (Levy & Fein, 1998).

Although labeling is regulated and is an important tool for consumers, the available space on the packaging is limited and some information cannot be reported even if they are important to consumers. Information overloading represents a potential source of danger to costumers, as it can deter them from making optimal decision.

In response to all the issues mentioned above, the Mobile Services for Agro-food (Ms4A) project was created to propose a state of the art solution to all these problems. The primary goal of this project is to facilitate consumers' access to immediate, complete and constantly updated information on the most various aspects of a product of interest. Through the use of a smart label technology, the basic idea is that more information could be stored in the product's package without incurring into information overloading problems.

Without going too deep into the technicalities of the smart labels, the relevant thing to say is that through this mature and well standardized technology, producers can now provide their customers with a far greater amount of information that is no longer limited to the size of the product's package. All this information can be easily accessed by any personal smartphone or any similar device with camera or RFID functionality.

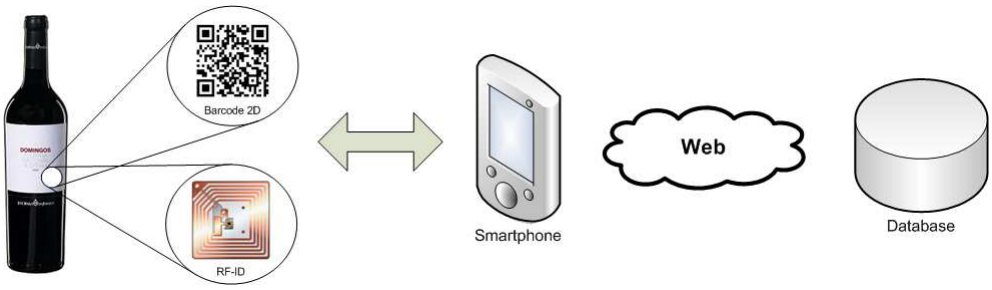

Figure 1 - Ms4A General Schema

Many studies have shown that costumer's individual profiles, like instruction, nutritional knowledge, income, and also age and gender can affect purchasing behavior (Nayga, 1996). In other words, this means that different costumer profiles value different kinds of information and therefore behave differently when purchasing.

The idea of increasing the available information without having overload issues can only be achieved if not all the information available is shown to the consumer. In order to provide only the relevant information for each costumer we need to have access this person's personal profile i.e. information such as, sex, occupation, age, income, instruction level, body mass index (BMI), intolerances, special diets, alimentary restrictions, etc. Only through the analysis of this information we are able to categorize the costumer and provide the appropriate information to each and every person. For more information on this topic refer to (Bantere & Cavaliere, 2011).

2.1.1. Ms4A potential benefits

The main innovation point in this project regards the customized information provided to the users based on their profiles. Nowadays, when we refer to a certain product label we see no differentiation on the information displayed. Most times, we have to go through a whole bunch of seamlessly useless information in order to find exactly what we are looking for. It's important to realize that through this customized approach, we no longer need to attain ourselves to the limited information set currently displayed in products packages. Through the use of ICT technologies (Smart Labels) it's possible to store new kinds of information that might also be useful to consumers and are not currently available.

By providing a client/producer oriented service, the Ms4A project tackles many issues that concern both entities. For instance, the value to final users can be justified as follows:

- Time saving: Information is easier to access and reflects the consumers' personal needs.

- Money saving: Additional information helps consumers to make more conscious acquisitions, avoiding purchases that don't fit the consumers' needs.
- Saves on alternative technologies: e.g. phone calls, consulting, website searches, etc.

In the other hand, for producers, the value can be justified as follows:

- Product differentiation by making available a series of more rich and interesting information from a commercial perspective.
- Increases quality and visibility of its production processes through integration of traceability system within the company.
- Reconstructs the geographical distribution of its products consumption and on this basis creates knowledge for targeted marketing.
- Dissemination of added-value services.

Once we have demonstrated the relevance of this project for consumers and producers, nothing makes more sense than to extend our efforts towards trying to better understand the feasibility aspects of this project and its potential economic value.

2.1.2. Reference Scenarios

In a focus group study conducted by the Dipartimento di Economia e Politica Agraria, Agroalimentare e Ambientale (DEPAAA), some really interesting analysis were made regarding the general population impressions on labeled information in the Ms4A context (Bantere & Cavaliere, 2011).

In the first step of the interview, based on a list of information not currently available on the labels, the consumers were asked to express their level of interest on each item of this list. Subsequently, was examined what were the motivations for the expressed interest by consumers. In this scenario, consumers responded the most to the following information:

1) Animal welfare (81%) - perceived as the result of a product with a higher quality.
2) Breeding (72%) - perceived as the result of a product with a higher quality.
3) Packaging (71%) – because of concerns related to the environmental impact caused by packaging materials.
4) Food Miles (67%) – perceived as important for products such as milk, water, eggs, fruit, and vegetables that deteriorate rapidly.

In the second step of the interview, based on a list of information that is generally available on the labels. Consumers were asked to express their level of interest on each item of this list. In this scenario, consumers responded the most to the following information:

1) Origin of the product (86%)
2) Presence or absence of GMO (72%)

3) Energy (70%)
4) Fat content (69%)
5) Organic agriculture (69%)
6) Absence of pesticides (67%)
7) Sugar content (67%)

The aim of the third step of the focus group was to investigate the interest of consumers in using the so called “smart label”. The main responses in this scenario were:

- Because of the easiness and the convenience of using smart labels, 81% of respondents affirm they would spend more time in getting information.
- Only 39% of respondents agree for a higher price in products for the smart label, while the remaining 61% wouldn’t like to pay any surplus.
- 72% of consumers agree to give private information in order to receive more personalized information. Only 28% of participants would not agree to give any private information.

By analyzing these reference scenarios, we can get a clearer understanding of what society’s needs are and what is the acceptance of this new solution among the general population.

2.2. Motivation

As mentioned in the previous section, the population increased interest in food knowledge, raises the question of the role played by food labels in fulfilling this new need. In this new context, issues such as space restriction, personalized information, lacking of information or data overload need to be addressed.

The IT field is evolving and developing every day. Enabling technologies such as mobile devices, internet, computers, databases, etc. are shaping the way that people communicates with one another, gets work done, and spends free time. In this constantly mutating environment, the choices we make need to be constantly rethought in order to adapt to the always changing population needs.

New technologies such as smart labels, which have been successfully implemented in warehousing applications, are now a new option for revolutionizing the way we label our products and the way we access this information. It’s important to understand that with this change of labeling paradigm, new issues arise regarding the new background integrated infrastructure that has to be built in order to support the new labeling features proposed. More information on this issue will be discussed in the next session.

2.3. Goals

What we have today, in many different areas, is a great quantity of relevant information divided and physically distributed throughout the web or as part of the business supply chain infrastructure. The idea of the Ms4A project is to identify and unify these heterogeneous sources of information in the agrofood business and by applying specific integration methodologies, optimize/maximize the information usage.

Despite the evidences that the Ms4A project represents a labeling solution of real value to customers and producers, just like any other IT project, before we implement any solution, the proposed scenarios of applications must be modeled/designed and evaluated through a qualitative/quantitative economic model in order to formally evaluate the convenience of this project.

This thesis goal can be divided in three parts:

1) Design/modeling of a data integration solution based on the Ms4A reference scenario presented in this chapter. The theoretical overview on how to accomplish this first step will be covered in chapters 3 and 4.
2) Application of the economic model to the integrated database solution in order to evaluate if there is indeed an actual increase in the information value, utility and quality. The theoretical overview of the economic model will be covered in chapter 4.
3) Cost variables examination in order to enable a cost/benefit analysis of the proposed data integration solution. For this part, the best approach would probably be a benchmark analysis of costs, based on previous data integration projects.

By the time all these steps are successfully accomplished, we'll hopefully have collected enough information in order to deliver a reasonable conclusion on the feasibility of the Mobile Services for Agrofood project.

3. STATE OF THE ART

3.1. Data Integration

Ever since the early 1960s the use of databases has become ever more prevalent in our modern societies. The storage of data has become a big part of our ability to use and manipulate information in order to generate knowledge. The increased use of databases and the development of collaboration between organizations have naturally led to the need for these organizations to make the data they own available to each other and also to their clients. In this way they are able to generate knowledge and in general make the information at their disposal more valuable and useful.

Data integration involves combining data residing in different sources and providing users with a unified view of these data (Lenzerini, 2002). This process becomes significant in a variety of situations, which include both commercial and scientific domains. In business circles, data integration is frequently referred as "Enterprise Information Integration" (EII).

Enterprise Information Integration, is a process of information integration, using data abstraction to provide a unified interface (known as uniform data access) for viewing all the data within an organization, and a single set of structures and naming conventions (known as uniform information representation) to represent this data; the goal of EII is to get a large set of heterogeneous data sources to appear to a user or system as a single, homogeneous data source (Halevy, 2005).

The issue of data integration generally arises when the component systems already exist, that is to say, integration is generally a bottom-up process. This topic has become ever more important in the modern day business environment as large corporations realize the importance of the data which is in their possession and in their information systems. Information that can be drawn from this data can give businesses a competitive advantage and help them survive highly competitive environments.

Another application of data integration is the combining of data sources that exist throughout the World Wide Web. The fact that web data sources are autonomous and heterogeneous makes the problem of web integration particularly tricky. Web integration can lead to the production of much more reliable search engines.

3.1.1. Data Integration Benefits and Applications

There is a manifold of applications that benefit from integrated information. For instance, in the area of business intelligence (BI), integrated information can be used for querying and reporting on business activities, for statistical analysis, online analytical processing (OLAP), and data mining in order to enable forecasting, decision making, enterprise-wide planning, and, in the end, to gain sustainable competitive advantages. For customer relationship management (CRM), integrated information on individual customers, business environment trends, and current sales can be used to improve customer services. Enterprise information portals (EIP) present integrated company information as personalized web sites and represent single information access points primarily for employees, but also for customers, business partners, and the public. Last, but not least, in the area of e-commerce and e-business, integrated information enables and facilitates business transactions and services over computer networks.

With typical investment in data integration tools falling in the range of $200,000 to $500,000 for software licensing and $50,000 to $100,000 for annual maintenance (Gartner, 2008), it's no surprise that so much investments is made either in the public or private sectors to foment researches in this field.

Now, if you are wondering why costs revolving data integration tools and maintenance are so high, the answer not surprisingly is the obvious one, implementing data integration is certainly not as easy as it sound. But let's not get ahead of ourselves! The issues regarding data integration will be covered in the next session, for now, let's focus on the reasons why organizations spend so much money in this solution.

As mentioned before, although not always simple, through data aggregation many positive things can be accomplished. Table 1 summarizes the main improvement points that can be achieved through data integration initiatives (Adelman & Moss, 2003).

Table 1 - Data Integration Benefits

Benefits	Comment
Availability / Accessibility	Asset data that is easily retrieved, viewed, queried, and analyzed by anyone within an agency.
Timeliness	Well-organized data can be quickly updated; one input will often apply the data across a variety of linked systems, and the information can be time-stamped to reflect its currency.
Accuracy and Integrity	Errors are greatly reduced because the integration environment drives a higher quality of input and can include automatic or convenient error checking and verification.
Consistency and Clarity	Integration requires clear and unique definition of various types of data, avoiding confusion or conflict in the meaning of terms and usage.
Completeness	All available information, including both historical and recent data, is accessible in an integrated database, with any missing records or fields identified and flagged via the integration process.
Reduced Duplication	Identical data is eliminated reducing the need for multiple updates and ensuring everyone is working from the exact same information.
Faster Processing and Turnaround Time	Less time is spent on consolidating and transmitting data to various users in the agency. The integrated data environment saves time by eliminating consolidation and transmittal to disparate users and allows many users to conduct separate analyses concurrently.
Lower Data Acquisition and Storage Cost	Data are collected or processed only once, and the information is consolidated and stored at locations supporting optimal convenience and ease of maintenance.
Informed and Defensible Decisions	Highly organized, comprehensive databases allow users to drill down through successive levels of detail for an asset, supplying more information to support decisions and supporting different types of analysis using various data combinations.
Enhanced Program Development	Comprehensive and coordinated system information advances program development by providing timely data for high-priority actions, promoting efficient distribution of funding among competing programs, and improving consistency in programs from year to year and across departments, among other benefits.
Greater Accountability	Data integration allows rapid and more accurate reporting of costs and accomplishments, including full attribution of results to relevant agency units and functions.
Lower Costs	Reducing the programs and the programmers that support the redundant systems. Despite the usually high cost of implementation, overall system maintenance costs drop considerably.

The integration of related data sources means that users are querying and manipulating a larger data set and for this reason they are able to retrieve more absolute and complete results thus favoring integrated decision making. Finally, the last aspect that needs to be taken in consideration is the system security issues. Studies have shown that data integration initiatives aid system security by focusing protection efforts in one centralized access point that can be more thoroughly and easily monitored.

3.1.2. Data Integration Issues

To achieve this holistic view, all data has to be represented using the same abstraction principles (unified global data model and unified semantics). This task includes detection and resolution of schema and data conflicts regarding structure and semantics.

First order of work is to resolve any inconsistencies between the schemas of the physical data sources. The structure and semantics of the individual schemas must be analyzed in order to identify naming inconsistencies, implicit attributes, absent attributes and other differences so that these may be rectified and a new data model developed. In general, however, there is no quick fix solution to integrating several data sources and to a large extent the approach taken, in each case, is governed by the individual characteristics of the information systems being integrated (Tsierkezos, 2010).

In general, information systems are not designed for integration. Thus, whenever integrated access to different source systems is desired, the sources and their data that do not fit together have to be coalesced by additional adaptation and reconciliation functionality.

While the goal is always to provide a homogeneous, unified view on data from different sources, the particular integration task may depend on (Dittrich & Jonscher, 1999)

- the architectural view of an information system,
- the content and functionality of the component systems,
- the kind of information that is managed by component systems (alphanumeric data, multimedia data; structured, semi-structured, unstructured data),
- requirements concerning autonomy of component systems,
- intended use of the integrated information system (read-only or write access),
- performance requirements, and
- the available resources (time, money, human resources, know-how, etc.)

Additionally, several kinds of heterogeneity typically have to be considered. This include differences in

- hardware and operating systems,
- data management software,
- data models, schemas, and data semantics,
- middleware,
- user interfaces, and
- business rules and integrity constraints.

As of 2010 some of the work in data integration research concerns the semantic integration problem. This problem addresses not the architectural structuring of the integrated solution, but how to resolve semantic conflicts between heterogeneous data sources. For example if two companies merge their databases, certain concepts and definitions in their respective schemas like "earnings" inevitably have different meanings. In one database it may

mean profits in dollars (a floating-point number), while in the other it might represent the number of sales (an integer). A common strategy for the resolution of such problems involves the use of ontologies which explicitly define schema terms and thus help to resolve semantic conflicts.

Finally, it's important to point out that most of the issues mentioned above are often times consequences of the fact that system developers rarely build application to be data-integration-friendly. Modern times require a mentality change in the IT industry. Application fragmentation begets data fragmentation. More often than not, applications are deployed as patches or quick fixes and in this practice developers rarely have the discernment of thinking through the consequences of their choices to the system as a whole in the future.

3.1.3. Data Integration Approaches

Information systems can be described using a layered architecture, as shown in Figure 2: On the topmost layer, users access data and services through various interfaces that run on top of different applications. Applications may use middleware — transaction processing (TP) monitors, message-oriented middleware (MOM), SQL-middleware, etc. — to access data via a data access layer. The data itself is managed by a data storage system. Usually, database management systems (DBMS) are used to combine the data access and storage layer.

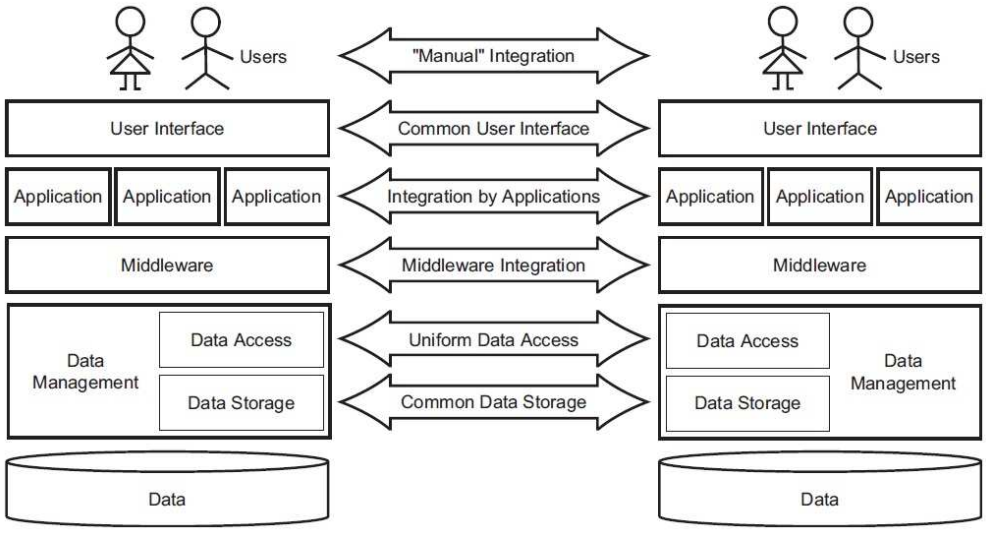

Figure 2 - General Integration Approaches on Different Architectural Levels

Seen through an architectural perspective, literature shows that there are several different ways to address the integration problem. The classification presented on Table 2 is based on (Dittrich & Jonscher, 1999) and distinguishes integration approaches according to the level of abstraction where integration is performed.

Table 2 - General Integration Approaches

Approach	Description
Manual Integration	Here, users directly interact with all relevant information systems and manually integrate selected data. That is, users have to deal with different user interfaces and query languages. Additionally, users need to have detailed knowledge on location, logical data representation, and data semantics.
Common User Interface	In this case, the user is supplied with a common user interface (e.g., a web browser) that provides a uniform look and feel. Data from relevant information systems is still separately presented so that homogenization and integration of data yet has to be done by the users (for instance, as in search engines).
Integration by Applications	This approach uses integration applications that access various data sources and return integrated results to the user. This solution is practical for a small number of component systems. However, applications become increasingly fat as the number of system interfaces and data formats to homogenize and integrate grows.
Integration by Middleware	Middleware provides reusable functionality that is generally used to solve dedicated aspects of the integration problem, e.g., as done by SQL-middleware. While applications are relieved from implementing common integration functionality, integration efforts are still needed in applications. Additionally, different middleware tools usually have to be combined to build integrated systems.
Uniform Data Access	In this case, a logical integration of data is accomplished at the data access level. Global applications are provided with a unified global view of physically distributed data, though only virtual data is available on this level. However, global provision of physically integrated data can be time-consuming since data access, homogenization, and integration have to be done at runtime.
Common Data Storage	Here, physical data integration is performed by transferring data to a new data storage; local sources can either be retired or remain operational. In general, physical data integration provides fast data access. However, if local data sources are retired, applications that access them have to be migrated to the new data storage as well. In case local data sources remain operational, periodical refreshing of the common data storage needs to be considered.

In practice, concrete integration solutions are realized based on the six general integration approaches presented. Table 3 lists and concisely comments some of most popular integration solutions out there (Ziegler & Dittrich, 2005).

Table 3 - Integration Solutions Examples

Solution	Description
Mediated Query Systems	Represent a uniform data access solution by providing a single point for read-only querying access to various data sources. A mediator that contains a global query processor is employed to send subqueries to local data sources; returned local query results are then combined (Wiederhold, 1992).
Portals	As another form of uniform data access, portals are personalized doorways to the internet or intranet where each user is provided with information tailored to his information needs. Usually, web mining is applied to determine user-profiles by click-stream analysis; that way, information the user might be interested in can be retrieved and presented.
Data Warehouses	Realize a common data storage approach to integration. Data from several operational sources (on-line transaction processing systems, OLTP) are extracted, transformed, and loaded (ETL) into a data warehouse. Then, analysis, such as online analytical processing (OLAP), can be performed on cubes of integrated and aggregated data.
Operational Data Stores	Second example of common data storage. Here, a "warehouse with fresh data" is built by immediately propagating updates in local data sources to the data store. Thus, up-to-date integrated data is available for decision support. Unlike in data warehouses, data is neither cleansed nor aggregated nor are data histories supported.
Federated Database Systems (FDBMS)	Achieve a uniform data access solution by logically integrating data from underlying local DBMS. Federated database systems are fully-fledged DBMS; that is, they implement their own data model, support global queries, global transactions, and global access control.
Workflow Management Systems (WFMS)	Allow to implement business processes where each single step is executed by a different application or user. Generally, WFMS support modeling, execution, and maintenance of processes that are comprised of interactions between applications and human users. WFMS represent an integration-by-application approach.
Integration by Web Services	Performs integration through software components (i.e., web services) that support machine-to-machine interaction over a network by XML-based messages that are conveyed by internet protocols. Depending on their offered integration functionality, web services either represent a uniform data access approach or a common data access interface for later manual or application-based integration.
Peer-to-peer (P2P)	Decentralized approach to integration between distributed, autonomous peers where data can be mutually shared and integrated. P2P integration constitutes, depending on the provided integration functionality, either a uniform data access approach or a data access interface for subsequent manual or application-based integration.

Among the many solutions presented above, the most popular and studied ones are probably Data Warehouses and Mediated Query Systems. For a better understanding of these two approaches, let's dig a little deeper and analyze more thoroughly what differentiate these two options.

The warehouse system extracts, transforms, and loads data from heterogeneous sources into a single common queriable schema so data becomes compatible with each other (see Figure 3). This approach offers a tightly coupled architecture because the data is already physically reconciled in a single repository at query-time, so it usually takes little time to resolve queries. However, problems arise with the "freshness" of data, which means information in warehouse is not always up-to-date. Therefore, when an original data source gets updated, the warehouse still retains outdated data and the ETL process needs re-execution for synchronization. For further information refer to (Kimball & Ross, 2002).

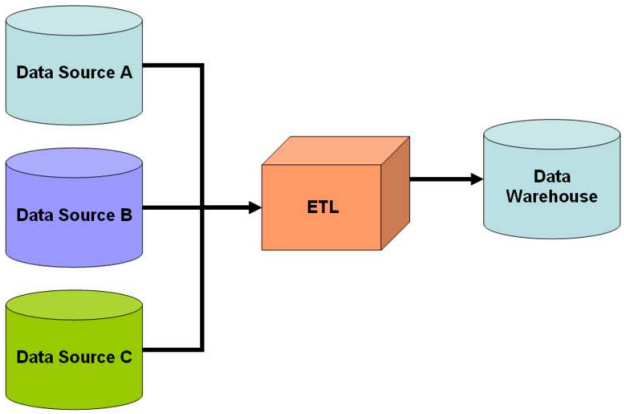

Figure 3 - Simple schematic for a data warehouse

As of recent years the trend in data integration has favored loosening the coupling between data and providing a unified query-interface to access real time data over a mediated schema (see Figure 4), which makes possible that information be retrieved directly from their original databases. This approach may need to specify mappings between the mediated schema and the schema of original sources, and transform a query into specialized queries to match the schema of the original databases. Therefore, this middleware architecture is also termed as "view-based query-answering" because each data source is represented as a view over the (nonexistent) mediated schema. For more information refer to (Domenig & Dittrich, 2000).

Irrespectively of the method used for the specification of the mapping between the global schema and the sources, one basic service provided by the data integration system is to answer queries posed in terms of the global schema. Given the architecture of the system, query processing in data integration requires a reformulation step: the query over the global schema has to be reformulated in terms of a set of queries over the sources (Lenzerini, 2002).

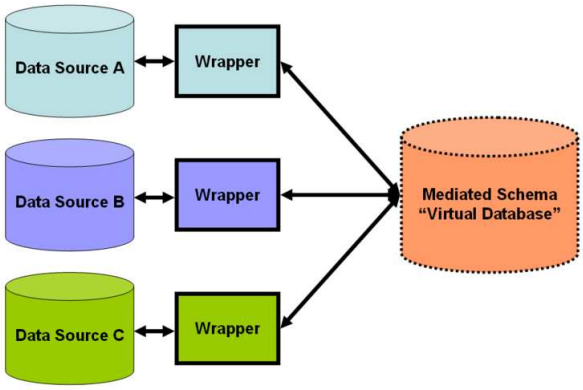

Figure 4 - Simple schematic for a view-based data-integration solution

The data integration systems we are interested in this work are characterized by an architecture based on a global schema and a set of sources. The sources contain the real data, while the global schema provides a reconciled, integrated, and virtual view of the underlying sources. Modeling the relation between the sources and the global schema is therefore a crucial aspect. Two basic approaches have been proposed for this purpose. The first approach, called global-as-view (GAV), requires that the global schema is expressed in terms of the data sources. The second approach, called local-as-view (LAV), requires the global schema to be specified independently from the sources, and the relationships between the global schema and the sources are established by defining every source as a view over the global schema. As it will surely come in handy, let's take a brief look at the differences between the LAV and GAV approaches, as well as their pros and cons.

3.1.4. Local-as-View vs. Global-as-View

Generally speaking, it is well known that processing queries in the LAV approach is a difficult task. Indeed, in this approach the only knowledge we have about the data in the global schema is through the views representing the sources, and such views provide only partial information about the data. Since the mapping associates to each source a view over the global schema, it is not immediate to infer how to use the sources in order to answer queries expressed over the global schema. On the other hand, query processing looks easier in the GAV approach, where we can take advantage that the mapping directly specifies which source queries corresponds to the elements of the global schema. Indeed, in most GAV systems, query answering is based on a simple unfolding strategy.

From the point of view of modeling the data integration system, the GAV approach provides a specification mechanism that has a more procedural flavor with respect to the LAV approach. Indeed, while in LAV the designer may concentrate on declaratively specifying the content of the source in terms of the global schema, in GAV, one is forced to specify how to get the data of the global schema by means of queries over the sources, thus making LAV a more reasonable choice for systems in which data sources constantly change.

In more reachable words, Table 4 summarizes the main differences between these two approaches.

Table 4 - Local-As-View vs. Global-As-View

Dimensions	LAV	GAV
Quality	Quality depends on how well we have characterized the sources.	Quality depends on how well we have compiled the sources into the global schema through the mapping.
Modularity	High modularity and reusability (if the global schema is well designed, when a source changes, only its definition is affected).	Whenever a source changes or a new one is added, the global schema needs to be reconsidered.
Query Processing	Query processing needs reasoning (query reformulation complex).	Query processing can be based on a simple unfolding strategy (query reformulation is easier).

3.2. Further Definitions

Without going too deep in each of the topics below, our intent in this section is to establish a basic vocabulary (based on literature) that might assist us during the explanation of this research. Also, especially in the model analysis, the concepts treated here will come in handy for a better understanding of the variables involved in data integration.

3.2.1. Data vs. Information

People often miss the subtle difference between data and information and use these two words interchangeably. In an informal environment not being able to differentiate these two concepts might not be frown upon, but when it come to information systems it's imperative to comprehend both definitions.

Data are plain facts. When data are processed, organized, structured or presented in a given context so as to make them useful, they are called Information. It is not enough to have data (such as statistics on the economy). Data in themselves are fairly useless. But when these data are interpreted and processed to determine its true meaning, they become useful and can be called Information (Bellinger, Castro, & Mills, 2004).

According to Russell Ackoff, a systems theorist and professor of organizational change, data, information and knowledge can be differentiated as shown below (Ackoff, 1989).

- **Data:** Often viewed as the lowest level of abstraction from which information and then knowledge are derived. Raw data, i.e. unprocessed data, refers to a collection of numbers, characters, images or other outputs from devices that collect information to convert physical quantities into symbols (Data on its own carries no meaning).
- **Information:** data that are processed to be useful, but does not have to be; provides answers to "who", "what", "where", and "when" questions. Information is data that has been given meaning by way of relational connection. In computer parlance, a relational database makes information from the data stored within it.
- **Knowledge:** application of data and information; answers "how" questions. Knowledge is the appropriate collection of information, such that its intent is to be useful.

When it comes to data integration, our final product is information. Information itself can be evaluated in many different levels namely quality, value, utility, capacity, cost, etc. As to interpret how information changes through data integration, it's important to have at least a basic understanding of all these variables. The next sections attempt to concisely define these concepts.

3.2.2. Information Variables

Information Quality

Information quality is a term to describe the quality of the content of information systems. It is often pragmatically defined as: "The fitness for use of the information provided."

Information quality is one of the measures of the value that information provides to the user. "Quality" is often perceived as subjective and the quality of information can then vary among users and among uses of the information. In an attempt to deal with this subjectivity, qualified professionals primarily representing the researchers' guild, have at one point or another identified particular metrics for information quality. They could also be described as 'quality traits' of information, since they're not so easily quantified, but rather subjectively identified on an individual basis (Wang & Strong, 1996). Table 5 lists the main information quality dimensions considered nowadays.

Table 5 - Information Quality Dimensions

Dimensions	Comments
Accuracy	Is the degree to which data correctly reflects the real world object or an event being described.
Accessibility	Accessible information is information that can be obtained when needed.
Completeness	Is the extent to which the expected attributes of data are provided. It is possible that data is not available, but it is still considered completed, as it meets the expectations of the user.
Consistency	Means that data across the enterprise should be in synch with each other (absence of any violation to business rules in a database.
Reliability	Refers to the accuracy and completeness of computer-processed data, given the intended purposes for use.
Timeliness	Timeliness refers to the currency of the information presented to the users. Currency of data or information is the time gap between the occurrences of an event until its presentation to the user.

Information Value

Information value can be defined as the amount a decision maker would be willing to pay for information prior to making a decision. The key component for information value analysis is to understand the relevance of the information provided to the user. In other words, does the information address its user's needs? If not, that user will find the information inadequate regardless of how well the information rates along other dimensions such as accuracy, accessibility, completeness, consistency, etc.

Even though information quality does not guaranty great information value, it's known for a fact that information value decreases if data contains errors, inconsistencies or out-of-date values. Therefore, high information quality levels can be considered as an initial parameter for the potential usefulness of the information objects but cannot assure it.

The correlation between information value and accuracy has been analyzed in (Moody & Walsh, 1999): the higher the accuracy of information, the higher its usefulness and value. Low accuracy levels can be very costly since they can cause both operational errors and incorrect decision-making. (Moody & Walsh, 1999) also point out that information value also depends on the age of the information.

Information is often very dynamic and its validity decreases over time. With the advent of the web, new quality dimensions have been investigated for characterizing the quality of an information source, e.g. believability, which considers whether a certain source provides data that can be regarded as true, real and credible, and reputation or trustworthiness that considers how trustable the information source is.

Information has become an element of commerce. In earlier times, success was based on such criteria as control of finance, physical resources, writing, food, fire or shelter. Today, successful people and businesses are those who control information: its development, access, analysis and presentation. We refer to our era as the "Information Age." We buy and sell information, sometimes with money and sometimes by trading it for other information. Information, as an element of commerce, is a commodity, yet there are no convergent, universally accepted accounting or economic theories of information (Fenner, 2002).

Information passes through many stages before it has value to anyone. It exists first in a latent state, waiting for the right paradigm or perspective, long before anyone recognizes it to be information. Then we realize that raw, unorganized data may be of some use. We collect it, organize it, analyze it and draw conclusions from it. Both the information and our conclusions can be communicated. Only when information has been comprehended, can we value it and respond to it (Howard R. , 1966).

A determination of the actual value of information can be made only at this final stage. Data has no value in itself; its value is derived from its understanding and subsequent application. Before this last stage we can do no more than estimate the value we expect it to have. Society values only the product, or result, of information.

Due to this subjectivity in determining information value, more often than not, information providers and information seekers struggle to get into an agreement on what the information is actually worth.

Information Utility

Underlying the laws of demand and supply is the concept of utility, which represents the advantage or fulfillment a person receives from consuming a good or service. Utility is an abstract concept rather than a concrete, observable quantity. The units to which we assign an “amount” of utility, therefore, are arbitrary, representing a relative value. Total utility is the aggregate sum of satisfaction or benefit that an individual gains from consuming a given amount of goods or services in an economy (Investopedia).

As mentioned before, society regards information as a commodity (marketable item produced to satisfy wants or needs) and the possession of it as an asset. Under these assumptions, information utility tries to measure the fitness or worth on a certain piece of information to some purpose or end. As seen in information quality and value, information utility is also a subjective concept which makes the attempt of quantifying it a laborious and imprecise activity.

Information Capacity

One predominant theme in much of this work is to build a new schema from existing ones using various structural manipulations. The new schema is intended to have equivalent information capacity with the original ones, or to subsume the information capacity of the original schemas in some sense.

The concept of information capacity investigated within data integration architectures, is seen as the increment in the number of queries that can be expressed over a set of databases integrated in a data integration architecture, and that could not be performed querying databases locally (Cappiello & Francalanci, 2011).

It's important to notice that the concept of capacity and quality seems to evoke an intrinsic property of data and information, a potential that can be defined and evaluated independently from the usage, while value and utility seem to evoke a property that depends on several factors, but especially the context and the usage.

Information Cost

Information can, to some extent, be valued and costed in the same way as the other assets of organizations, and included in their financial reports. As inventory, information goes through the value-added stages of raw material (events or processes to be measured), work-in-progress (information in development), and finished goods (marketable information). Information gathering and presentation require capital investment and human labor. Besides being costly to acquire, information incurs management costs. Like physical assets, information faces quality control inspection before it can be distributed. Information is subject to just-in-time requirements, just like physical inventory. Left on its own, its value may depreciate over time (Fenner, 2002).

The adoption of data Integration architectures can be designed as part of a corresponding IT project and, thus, involves different types of costs, namely design time costs and run time costs (Cappiello & Francalanci, 2011). Design time costs are:

- **Source wrapping cost:** Local sources have to be registered in the data architecture and wrapped to enable information extraction. This cost depends on the number of sources that have to be integrated.
- **Mapping cost:** Design cost for the definition of the relations between global schema and local schemas.
- **Maintenance cost:** Maintenance cost of the integrated architecture. This is composed of the maintenance costs for wrapping and for mapping operations. Maintenance costs also vary with the number of sources integrated.

- **Data quality cost:** Costs associated with data quality assessment and cleaning operations.

Run time costs concern with query execution costs and depend on the frequency with which the information is queried and updated.

All the costs mentioned above will be more thoroughly covered in the methodology chapter of this research. For now, last thing worth mentioning is that just like in any IT project, technology mediating the information provision and its use/consumption cannot be taken for granted, and hence it has to be included in the cost analysis. Added to all design costs mentioned above, data integration also requires hardware and software investment (licensing) that account for fixed costs that must be considered in the beginning of any integration project (Howard P. , 2010).

Information Structure

In computer science, a data structure is a particular way of storing and organizing data in a computer so that it can be used efficiently (Black, 2004). Different kinds of data structures are suited to different kinds of applications, and some are highly specialized to specific tasks. For example, B-trees are particularly well-suited for implementation of databases, while compiler implementations usually use hash tables to look up identifiers.

Data structures are used in almost every program or software system. Data structures provide a means to manage huge amounts of data efficiently, such as large databases and internet indexing services. Usually, efficient data structures are a key to designing efficient algorithms. Some formal design methods and programming languages emphasize data structures, rather than algorithms, as the key organizing factor in software design.

In data integration, information structure is a very relevant issue especially when it comes to the design of the integrated databases. Understanding Local sources' structure and defining the more convenient data structure for a mediated schema is imperative to the efficiency and easiness of how the data is manipulated.

4. METHODOLOGY: ASSESSING THE VALUE OF INFORMATION INTEGRATION

The main goal of this chapter is to set a basic understanding of the tools, theories, methods, assumptions and above all, the procedures that will be carried out throughout this study in order to arrive at its results and justify its believability. By having a good understanding of the concepts explained in this part, anyone should be able to replicate this study or properly apply it in different scenarios.

Section 3.1 briefly describes the steps taken in the development of this project, providing us a quick glimpse of where we are, where we want to go and what we need to do to get there. Section 3.2 and 3.3 are probably the most significant parts of this chapter, and they describe thoroughly the model in which we base ourselves in order to structure, analyze and evaluate all the issues previously raised on data integration. Section 3.4 gives us a short didactical example of a schema integration methodology that will later be extended and applied to the agrofood scenarios. Finally, Section 3.5 focuses on finally connecting the dots of everything that has been presented so far. How do all the concepts presented lineup and gives us enough resources to study case the agrofood business.

4.1. Project Roadmap

As previously mentioned, this section aims at listing the activities involved in this project development. Although things not always take place linearly, the following list can be considered as a guideline of how things progressed in this research. As these steps have been proven effective and reasonably efficient in providing positive results, hopefully by following these same milestones, the results achieved can be easily replicated. The project had the following phases:

1) Problem analysis (Ms4A): Just like any regular project, the first step towards solving a problem is to understand what we are dealing with. To do so, a couple questions have to be answered:
    a. What is the industry and how does it operate?
    b. Who are the stakeholders? (Producers, Clients)
    c. What are the issues? (Technological, social)
    d. What are the goals?

2) Literature study (Enabling technologies - Integrated databases): Once the problem is understood it's time to build knowledge to evaluate the best options for solving it.
3) Model decision and study: Choose the more appropriate model and gain a better understanding of it.
4) Scenarios creation and analysis: Suitable scenarios must be generated based on the industry analysis and the model's requirements.
5) Model application to the identified scenarios: Once the scenarios have been identified it's time to apply the model and seek results.
6) Result analysis
7) Conclusion

As can be easily seen, the approach taken is pretty straight forward and follows accordingly all the scientific methodology standards.

4.2. A model of information capacity and value of integration technologies

When it comes to understanding how integration technologies affect data and information, it's important to focus on two basic concepts that have been often associated with these two elements, namely capacity and value. Not surprisingly, innumerous studies have been conducted in order to answer the following research questions:

- How should the concept of information capacity be modeled?
- What is the maximal total increase in information capacity achieved using data integration technologies over a set of databases?
- How is information capacity related to quality of data, such as accuracy and completeness, and to the value produced by data integration initiatives?

In this section we intend to present a unified quality based model that will help us understand, quantify and validate how integration technologies influence data/information capacity/value. By later applying this model to an actual practical case (ms4A) we hope to gain a better understanding of relevant business strategic factors such as the utility, the value and the economic viability of the project.

The model here demonstrated is based on the article "Information value and information capacity: a unified quality based model" (Cappiello & Francalanci, 2011).

Literature analysis shows that information value is the most investigated concept among the ones previously introduced. Furthermore, different approaches to information value have been followed, depending on the disciplinary area and focus. In general, economy and knowledge oriented perspectives pay little attention to the type of technology enabling the

information provision and its use/consumption, whereas in management of information systems approaches to information value can be hardly disentangled from the technological and organizational resources and environment. As a consequence, a common interpretative framework for information value has to consider a set of characteristics which can provide dimensions suitable to evaluate information facets at different level of abstraction. As to technology, it is worth noticing that it represents an independent variable or weight for each characteristic. In the following paragraphs we are going to analyze a model for characterizing the information capacity of a technological data architecture, namely a set of databases federated through data integration (DI) technologies, and considering also several quality dimensions of data and information.

Figure 5 shows the proposed unified model where information value is related to the multiple factors namely information quality, information structure, information diffusion, and information utility (the current version of the model does not consider information risk). The main hypothesis underlying the proposed model is that information value has to be interpreted in terms of information utility, which is characterized by both the overall information capacity of the considered information systems and the costs related to the potential information capacity resulting from integration initiatives.

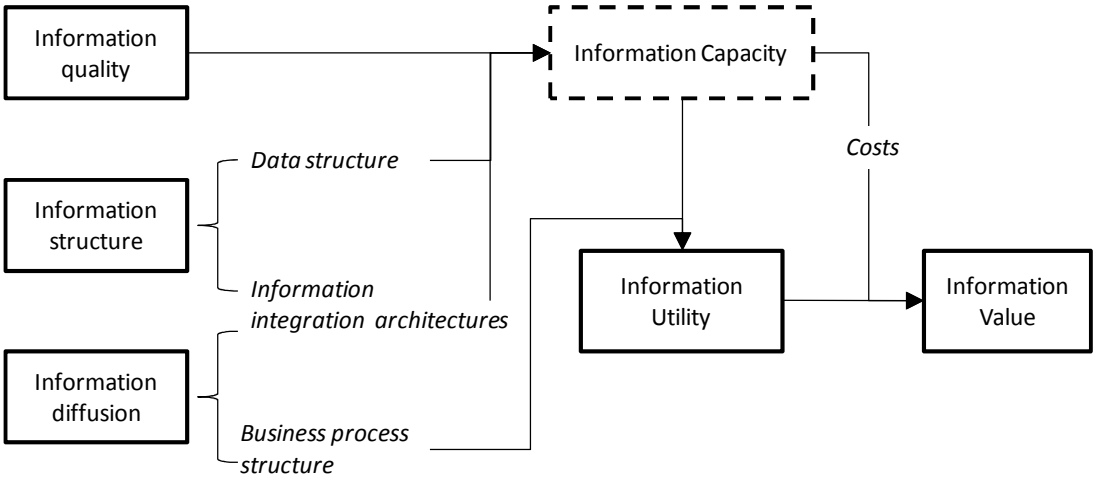

Figure 5 - The proposed unified model of information capacity and value

It is worth noticing the relevance we attribute to information capacity. Indeed, in the proposed model information capacity is a comprehensive information value characteristic, which is composed by i) information quality dimensions, ii) information structure (as the degree of integration of the data structure considered), iii) information diffusion (as integration architectures and technologies adopted).

4.3. Formalization of information capacity, utility and value

In order to formalize the model of Figure 5, we first provide a definition of a DI database architecture. We assume that databases in the architecture make use of the relational model. A data integration (DI) architecture is usually defined as a triple < *G, S, M* > where *G* is the global schema, *S* is a set of local schemas and *M* is the mapping between *G* and *S*. The global schema *G* can be defined as a set of entities $E = \{e_1, e_2, \ldots, e_j\}$ each one expressed in terms of a view on the set of local schemas $S = \{s_1, s_2, \ldots, s_m\}$; this type of mapping is called the Global as View (GAV) mapping, whereas in the Local as View mapping the entities of local schemas are expressed as views on the global schema (Lenzerini, 2002). In our approach, we assume that the global schema contains all the attributes of the local sources. The mapping *M* can be expressed in different ways depending on the architecture that is adopted, but in practice it defines, for each attribute of the global schema, its relationship with each attribute of each local schema.

In our approach we consider a set of data bases $DA_{DI} = [db_1, db_2, \ldots, db_m]$ integrated through a DI architecture and associated with the corresponding schemas $S = [s_1, s_2, \ldots, s_m]$. In the following, to formalize the framework, we use a graph-based approach.

We assume that each schema $s_i \in S$ can be represented as an oriented graph $s_i = \{N_i, R_i\}$, where:

- N_i is the set of nodes representing the tables included in the relational schema.
- R_i is the set of directed edges, representing foreign keys between tables. Each $r \in R$ can be represented as $r=(n_k, n_z)$ where n_k is the table that contains a foreign key that matches the primary key of the table n_z.

Each schema $s_i \in S$ is characterized by an application load AL_i that specifies the set of queries usually applied to the schema. Thus, AL_i can be defined as $Q_i = \{q_{i1}, q_{i2}, \ldots, q_{ih}\}$. In our model a query is expressed as a sequence of tables linked by using equijoin operations enabled by foreign keys. Using the graph representation, it is possible to state that a query q_{ih} is represented as a *path* between nodes of the oriented graph. We suppose that the queries have a maximum length L. In the definition of the global schema, the subgraphs related to the different local sources may have pair wise common overlapping nodes (tables). Such nodes are called *borderline nodes*. For each local source s_i, we call $BN_i \subseteq N_i$ the set of borderline nodes in s_i. For any schema s_i, for any borderline node $n_j \in BN_i$ we associate to n_j the set of queries BQ_{ij} of the application load AL_i that contain the borderline node n_j.

In the following we need to define the new set of queries NQ that can be expressed in the DI architecture and that could not be expressed on local schemas. In Figure 6 (a), we show examples of queries in NQ represented with dotted lines in a three sources architecture, and resulting from an extension of a query in the application load AL_1, represented with solid line. In Figure 6 (b) we provide an algorithm to evaluate NQ.

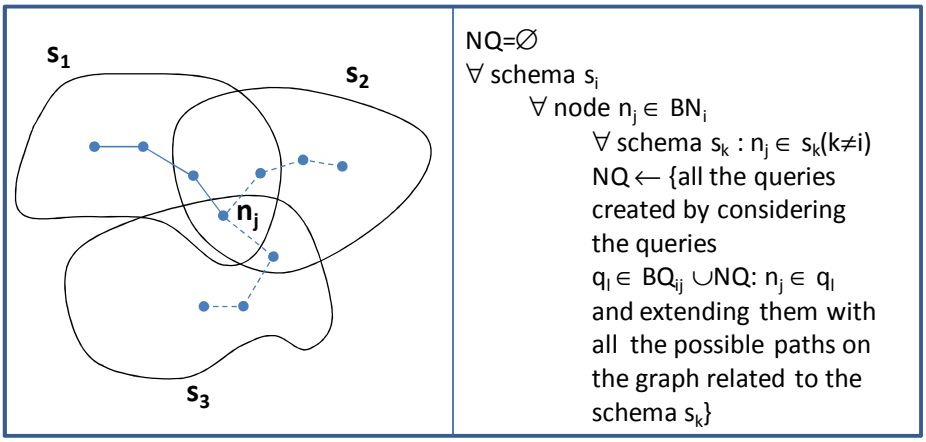

Figure 6 - Example and algorithm for the definition of the set of new queries NQ

We now proceed to define the concept of *information capacity*. We propose two different definitions of information capacity:

1. Intensional Information Capacity (IIC) that makes reference to schema and calculates the improvements in terms of new queries that it is possible to express on the DI architecture and
2. Extensional Information Capacity (EIC) that makes reference to database instances and measures the effect of data quality errors on the tuples that are linked in equijoins relating tables on different schemas (see later).

As a consequence of the above definitions and algorithms, the *intensional information capacity* IIC(DA_{DI}) of a data architecture is described by the following formula:

$$\mathrm{IIC}(\mathrm{DA}_{\mathrm{DI}}) = \frac{|NQ| + \sum_i |AL_i|}{\sum_i |AL_i|}$$

As to the *extensional information capacity*, we have to consider that in real life databases, data are characterized by errors of various types, and the number of retrieved instances is influenced by the quality of data. As an example, assume we have two databases made both of a unique table DB1 = [T1(A,B)] and DB2 = [T1(B,C)] and assume that the accuracy of values of B in DB2 is 0.9, meaning that 90% of values of B are correct. Furthermore, assume all the values of B in DB1 are correct (accuracy=1). In this case a join will not link the 100% of the tuples, but it will link only a subset depending on the distribution of values of B in DB1. The extensional information capacity (EIC) measures the ratio between the total set of the tuples I_k retrieved by all the queries $q_k \in NQ$, and the tuples I_k' retrieved in the ideal case where no data quality errors occur. Formally:

$$\mathrm{EIC}(DA_{DI}) = \frac{\sum_k |I_k|}{\sum |I_k^{'}|}$$

In our model we consider one data quality dimension, the *key accuracy*. Key accuracy of a table, in relational terms, can be measured as the percentage of key values that are contained in a look-up table of all valid values.

It is possible to improve the EIC by applying methods for the evaluation and improvement of the quality of instances see (Batini, Cappiello, Francalanci, & Maurino, 2009) and (Batini & Scannapieco, 2006). We define therefore a *quality improved extensional information capacity* as the ratio between the number of *correct* tuples retrieved and the number of all instances in the ideal case. Formally:

$$\mathrm{EIC}(DA_{DI}) = \frac{\sum |I_k| + \sum |I_k^c|}{\sum |I_k^{'}|}$$

Where I_k^c represents new instances discovered thanks to the data quality improvement techniques applied to the data integration architecture.

With the final goal of defining the value, we move now to the concept of *utility of a DI architecture*. In an organization, the activities are performed by means of business processes. A business process bp_w is a set of tasks that accomplish a specific organizational goal. Business processes are often supported by the information gathered from the databases. Thus, each query has the potential to produce useful information to the processes that submit it. In general, we can assume that, considering a generic query q_{ih} defined on the local sources s_i, a function U = util(q_{ih}, bp_w) exists, where U is an adimensional value that defines the importance of the query q_{ih} for the process bp_w. For an organization, the utility achieved by using the different local sources (s_1….s_m) can be expressed as:

$$Utility(s_1 \ldots s_m) = \sum_w \sum_i \sum_h U(bp_w, q_{ih})$$

With the adoption of an integrated architecture, each process has the capability to gather data from the integrated architecture using additional queries $q_k \in$ NQ. Considering that each query submitted by a process is associated with a specific utility, the DI architecture provides an additional utility *Utility(DA_{DI})* defined as:

$$Utility(DA_{DI}) = \frac{\sum_w \sum_k U(q_k, bp_w) + \sum_w \sum_i \sum_h U(bp_w, q_{ih})}{\sum_w \sum_i \sum_h U(bp_w, q_{ih})} \qquad q_k \in NQ$$

Similarly we could define the utility enabled by the extensional information capacity.

As an example of utility that a DI architecture can provide to a process, consider the case of a registry of households that is integrated with a customer registry; in this case a targeted marketing campaign can be enhanced with cross-sellings initiatives.

We now discuss how the concept of utility differs from the concept of value. To do so, we consider costs. The adoption of a DI architecture can be designed as a part of a corresponding IT project and, thus, involves different types of costs, namely design time costs and run time costs. Design time costs are:

- C_{wrap}: design cost for source wrapping. Local sources have to be registered in the data architecture and wrapped to enable information extraction. This cost depends on the number of sources that have to be integrated.
- C_{map}: design cost for the definition of the global schema and for mapping local schemas to the corresponding entities e_j of the global schema. This information is stored linking local entities and their corresponding global entity. In the following we focus on the design-time costs associated with constructing the mapping table. For each entity e_j, we model this cost as a function of the number of distinct queries involving e_j. For each query involving e_j, designers must understand the instance set involved in answering the query and store the corresponding information in the mapping table, accordingly. These costs are expressed as:

$$C_{map}(e_j) = d \cdot | Q_j |$$

where d represents the average cost of analyzing a query and storing the corresponding information in the mapping table, while $| Q_j |$ indicates the number of queries involving entity e_j.

Thus, the total mapping cost for DI architecture is:

$$C_{map} = \sum_j C_{map}(e_j)$$

- C_{maint}: maintenance cost of the integrated architecture. This is composed of the maintenance costs for wrapping and for mapping operations. Maintenance costs vary with the number of sources and the time elapsed from initial design.
- C_{DQ}: costs associated with data quality assessment and cleaning operations.

The total design cost is defined as:

$$C_{des} = C_{wrap} + C_{map} + C_{main} + C_{DQ}$$

Concerning run time costs, each entity e_j involves a runtime cost, c_{exe}, that depends on the frequency with which the entity is queried and updated, called respectively $cq(e_j)$ and $cu(e_j)$:

$$C_{exe}(e_j) = p_j \cdot cq(e_j) + r_j \cdot cu(e_j)$$

where p_j and r_j represent the average cost of querying and updating entity e_j.

The total runtime cost of the global schema E is the sum of the runtime cost of all entities $e_j \in E$ that is:

$$C_{exe}(E) = \sum_j C_{exe}(e_j)$$

Information value is defined based on the information utility and calculated as follows:

$$Value(DA_{DI}) = \frac{\sum_w \sum_k U(q_k, bp_w) - C_{tot}}{C_{tot}} \qquad q_k \in NQ$$

Where C_{tot} represents the sum of design cost (C_{des}) and runtime costs over a given time period, typically the expected lifetime of the architecture.

As it can clearly be seen, the proposed model provides a unifying framework for several paradigmatic concepts associated to information in the information systems literature, such as, besides capacity and value, the issues of information quality, structure, diffusion, cost and utility.

Once we have settled the theoretical grounds for our study, it might be convenient, before we jump into the more pragmatic part of this project, to see a more didactical and hand-on example of data integration. The following section will briefly introduce a simplified example of how data integration can be achieved as well as some of the issues that we might run into, and how to deal with them.

4.4. Schema Integration Methodology Example

In order to introduce the reader to the main features and problems of schema integration, we present the following example that was extracted from the article “A comparative Analysis of Methodologies for Data Schema Integration” (Batini & Lenzerini, 1986). In the hope of sparing us time with long explanations about why and how we are doing things in the following chapters, this example gives us a general overview and a solid starting point to analyze more complex and extensive cases (Ms4A).

The methodology described here is one among several methodologies available for schema integration. The figure below shows two different, basic conceptual schemas (Figure 7). The first one describes a dataset of books; the second one describes a dataset of publications. It must be noted that “Topics” in the first schema holds the same meaning as “Key Words” in the second. Also, it’s important to point out that “Publication” in the second schema is a more abstract concept than “Book” in the first schema. That is, “Publication” includes additional things such as proceedings, journals, monographs, etc.

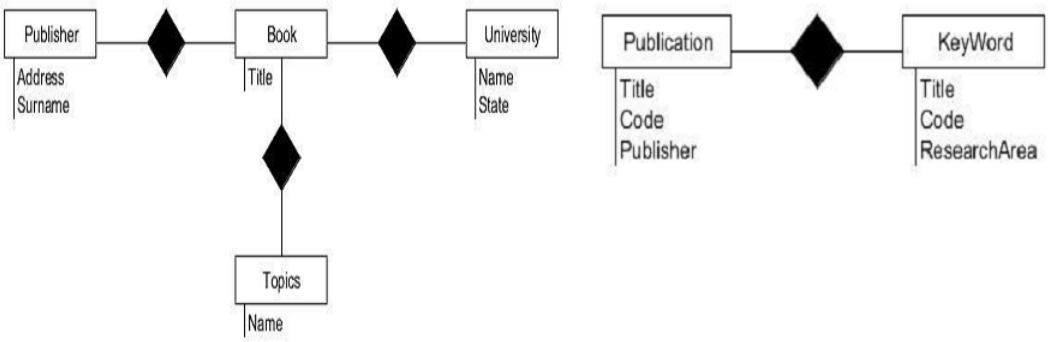

Figure 7 - Original Schemas

The following figures display the possible chain of events undertaken in order to integrate these two schemas. Firstly, since ”Key Words” and “Topics” hold the same meaning, i.e. represent the same real world object, in order to merge them we need to establish a common name for this entities. Either names work fine, for our convenience, let’s choose “Topics”. Figure 8 shows the new schema convention. Another issue that needs to be tackled is the fact that “Publisher” is an entity in the first schema but an attribute in the second one. In order to fix this structural incoherence we create a new “Publisher” entity in the second schema and add a new attribute, Name, to it (Figure 9). The next step in the integration process is to superimpose the two schemas as illustrated in Figure 10. Within the superimposed schema we notice that there is a subset relationship between “Book” and “Publication”. Figure 11 illustrates this subset relationship. Finally, we can simplify the representation by removing the properties in “Book” that are common to “Publication”. This is possible since the subset relationship implies that all the properties of “Publications” are implicitly inherited by “Book”. The final merged schema is shown in Figure 12.

Figure 8 - Renaming Key Words to Topics

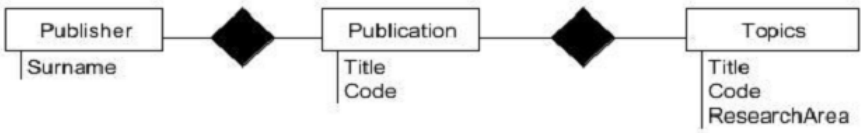

Figure 9 - Make publisher into an entity from an attribute

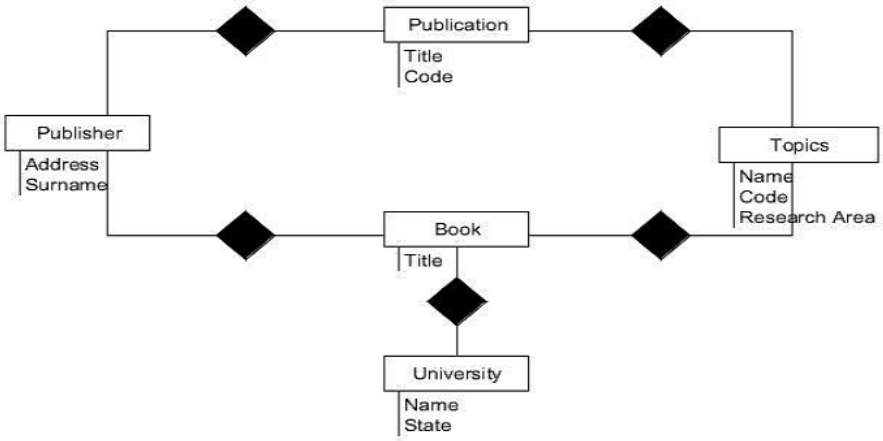

Figure 10 - Superimpose the two schemas

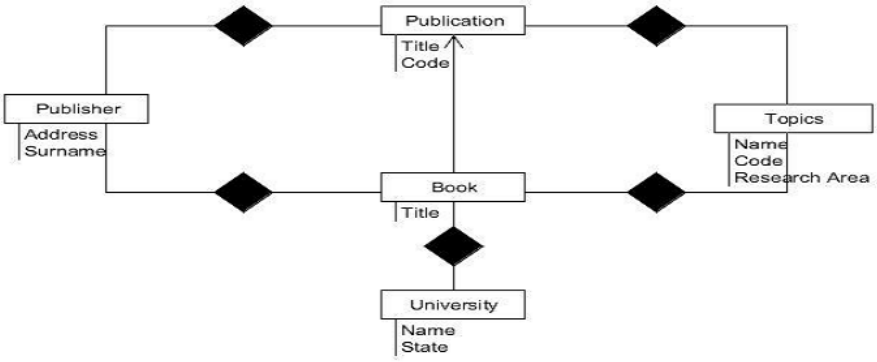

Figure 11 - Make Book a subset of Publication

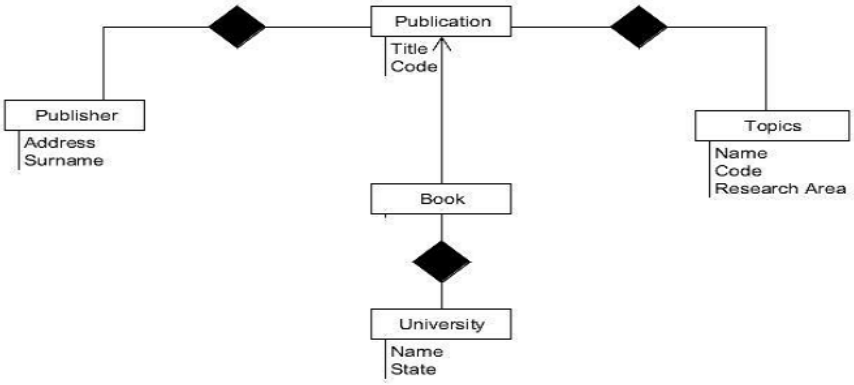

Figure 12 - Remove Books common properties

The above schema integration example is simple, but highlights some of the basic problems that system designers might have to contend with. It is amply evident from this

example that the integration of realistic sized component schemas can be a complex endeavor. The basic problems to be dealt with during integration come from structural and semantical diversities of schemas to be merged. For more information on the various causes for schema diversity and other integration methodologies, refer to (Batini & Lenzerini, 1986).

4.5. Data Integration and Mose for Agrofood

In order to finalize the methodology part of this study, this section intends to finally connect the dots between all the concepts that have been presented so far and the Ms4A project. Up to this point, the main focus of our work was to gather and analyze relevant information and tools on data integration. By mastering these ideas, we are now finally capable of applying these concepts in the Mose for Agrofood context, giving us a better understanding of the feasibility, the cost and the benefits of this project.

As we've seen on chapter 3, among the many approaches introduced for data integration, the two most accepted ones are Data Warehouses and Mediated Query Systems. Practice shows that Data Warehouses are a really reasonable choice when it comes to big corporations where all local data sources involved in the integration process are internally owned and managed. In our case, although we rely on supply chain local sources we also rely on multiple distributed internet based sources. Since these last ones are constantly subject to unpredictable changes, we face the fact that the wisest choice in this case is most probably the View-based data integration approach. Through this approach we gain in dynamism, and changes on the local sources can be more easily managed without critically affecting the system's performance.

Once we have decided on the View-based approach, we still have to go a little bit deeper and decide among the more specific mechanisms proposed for this purpose, namely global-as-view and local-as-view approaches. The pros and cons of these two techniques were already discussed in the previous chapter, and after reflecting on the appropriateness of each method in the Ms4A context, we decided for the global-as-view approach. The strongest argument that led us to picking this option was the fact that the target users of this project are, most commonly, ordinary people that don't have much background knowledge on information technology. For this reason, query processing should be as intuitive and transparent as possible. Even though this approach might increase considerably the designing efforts necessary for integrating the local data sources (especially when it comes to web sources), in the front-end we can offer a much more user-friendly interface that surely will appeals a lot more to the users.

Since web sources are generally autonomous, in many real-world applications (as in ours) the problem of mutually inconsistent data sources arises. In practice, this problem is generally

dealt with by means of suitable transformation and cleaning procedures applied to data retrieved from these sources.

Also, one of the main tasks in the design of a data integration system is to establish the mapping between the sources and the global schema, and such mapping should be suitably taken into account when formalizing a data integration system. Intuitively, the source schema describes the structure of the sources, where the real data are, while the global schema provides a reconciled, integrated, and virtual view of the underlying sources. The assertions in the mapping establish the connections between the elements of the global schema and those of the source schema.

At this point, probably the most important thing to be understood is that in any IT project before we start investing enormous amounts of money on hardware and implementation, it's of the highest priority to evaluate the values and costs involved in this endeavor. Having straightened out these facts, that's when the model previously presented comes in handy. Before we engage in this enormous project that involves integrating multiple distributed data sources we must first analyze if the investment we are about to make is actually going to fulfill our expectations and needs. By analyzing parameters such as value, utility, capacity and quality of information in an integrated solution, we are able to gather relevant information that will assist us in evaluating the possible outcomes of the project, thus allowing us to decide, with more concrete information, whether the approach we are following is indeed the more convenient one.

Now that we have bridged the gap between the studied theory and the project context, it's time to see how all this applies in the Mose for Agrofood scenarios.

5. EMPIRICAL VALIDATION: A CASE STUDY IN THE AGROFOOD DOMAIN

The objective of this chapter is to describe the application of the model of information capacity and value in the Mose for agrofood scenarios. By going step by step in the design process of the global schema we hope to apply all the data integration knowledge acquired so far in this study. Ultimately, by thoroughly analyzing factors such as information value, capacity, quality, utility and cost we should be able to arrive to a compelling conclusion on the feasibility of this project.

5.1. Global Schema Design

Once we have analyzed and understood how concepts such as information capacity, value, utility and quality can be unified and applied to integrated databases adopting data integration technologies, it's time to observe in a more pragmatic point of view how the model we previously studied can be applied to Ms4A scenarios.

When we look closely at the agrofood sector we notice that the main sources of information are either related to the product's supply chain or linked to the web. Obviously the information provided by these two sources are substantially different and complementary, therefore it's indeed important to consider both references.

Before we start looking into local sources schemas and global schema designs, let's take a brief look at what kind of information can be extracted from each of these sources:

Without paying too much attention to any traceability issues that the supply chain might present, relevant information in the supply chain can be extracted primarily from five different levels, namely the agricultural phase, transportation phase, industrial phase, distribution phase and finally from the consumption phase. Indeed, the supply chain in the food industry can be very heterogeneous depending on the specific type of food. These characteristics vary due both to the specific features of the considered food, but also due to the economic and industrial environment which is proper of a specific region, and to specific laws that regulate each given market. Table 6 summarizes some of the information that can be retrieved from the supply chain on each phase.

Table 6 - Supply Chain Phases vs. Information retrieved

Phase	Information
Agricultural	Sowing period
	Plot of land cultivated
	Farming techniques (use of pesticides ...)
	Harvest period
Transportation	Storage / Conservation
	Transportation conditions
	Kilometers traveled
Industrial	Control of requirements compliance upon receipt
	Conditions of storage and preservation
	Information on the transformation process
	Control of requirements compliance
	Packaging method
Distribution	Periodic sampling
	Conditions of storage and preservation
	Quality control upon receipt
Consumption	Nutritional Information
	Product origin information
	Process information

Not surprisingly, when it comes to information on agrofood products available on the web, we find ourselves overwhelmed with the enormous amount and variety of information that is out there. More often than not, the same kind of information can be found in hundreds of different sources, still also not rarely this many sources present conflicting information. Problems with data quality are spread throughout the web, especially when it comes to dimensions like accuracy, consistency, completeness and reliability. For this reason, despite the amazing amount of information that can be found on the web, we need to be very cautious when using this information since the collaborative environment that characterizes the internet makes it very difficult for users to trace back references of the data provided. Just to exemplify some of the information that can be retrieved from the web we have:

- Restaurants
- Recipes
- Nutritional Claims
- Health Claims
- Ingredients
- Sentiment Analysis

Figure 13 shows a summarized overview of the different sources from which information can be retrieved in the agrofood sector, as well as a small sample of the kinds of information found in each of these data sources.

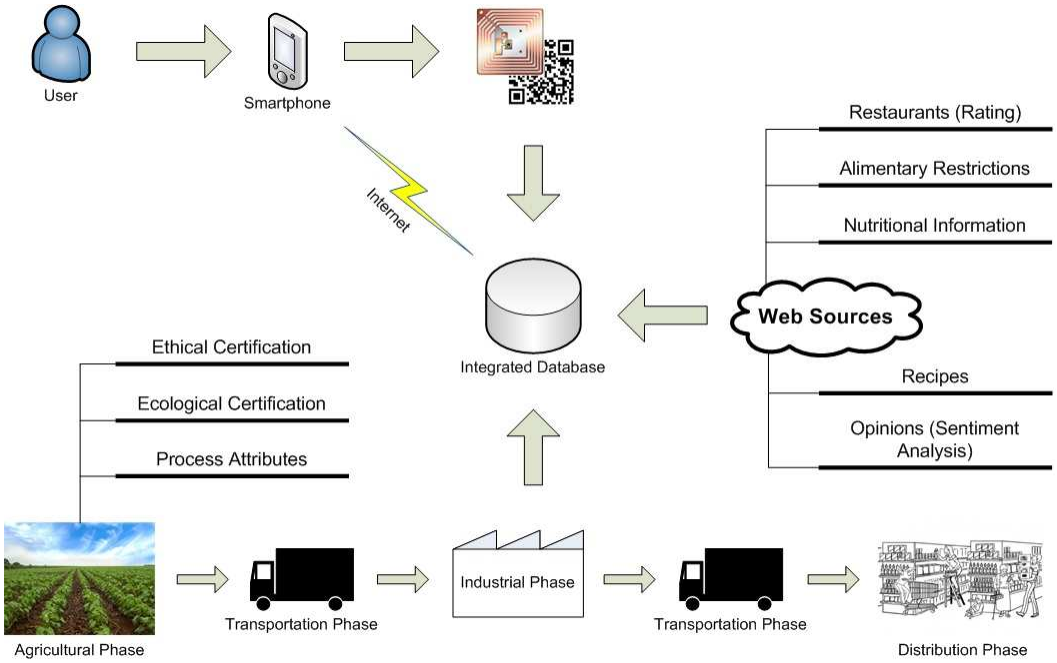

Figure 13 – Source of information in the agrofood sector

One of the most important aspects in the design of a data integration system is the specification of the correspondence between data in the local sources and their representation in the global schema. Such correspondence is modeled through the notion of mapping, as introduced in the previous chapters. The source mapping will determine how the queries posed to the system are answered.

As pointed out in the model description, the process of data integration is basically composed by 3 main steps, namely data cleaning, data wrapping and data mapping. The first one corresponds to quality assessment and cleaning operations of local sources. The second one regards the process of registering the local sources in the data architecture and wrapping them in order to enable information extraction. Finally, the last one corresponds to the efforts of defining the global schema, i.e. building the relationships between local schemas and the corresponding entities in the global schema.

Since our approach in this study is of a higher level of abstraction, for now, we are not going to focus much of our attention in the data cleaning process. The main reason for that is that in order to exemplify and justify data cleaning operations we would require pertinent local data sources with actual data as to be able to analyze its data quality. For further information regarding this subject refer to (Batini & Scannapieco, 2006) and (Batini, Cappiello, Francalanci, & Maurino, 2009)

The following examples demonstrate the process of source wrapping in the Ms4A scope. Although at some level these examples might feel a little bit idealized, if we keep in mind all the unavoidable data quality issues that we might face, they still hold great relevance for a

qualitative analysis of this project's information value, capacity, utility and quality, as well as it helps us estimate this project's economical viability.

- **Nutritional Claims:**

A product's nutritional information can generally be extracted either from the supply chain or the web. A famous example of this kind of internet source is the American website CalorieKing (CalorieKing Wellness Solutions, 2011). With an enormous database that includes a huge variety of dishes or individual ingredients this website provides free of charge information on food nutritional claims. The searches can be done by products/ingredients generic names or brands. Since the website focused on encouraging diet programs, it also features a lot of information on this topic as well as interesting related statistics on the food you are analyzing. Figure 14 shows a basic description of the relevant nutritional information of agrofood products that might be relevant to costumers (e.g. calories, carbohydrates, fat, proteins, fibers, minerals, vitamins, etc.).

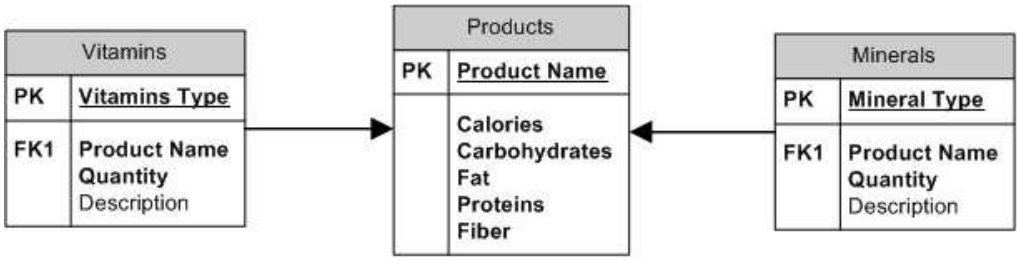

Figure 14 - Agrofood products nutritional claims

- **Ingredients/Components:**

A shelf product can have its ingredient information extracted either from the supply chain or the web. For reliability reasons and because of regulamentation issues we know for a fact that companies must have this kind of information in order to commercialize their products, therefore makes a lot more sense to rely on the supply chain databases in order to retrieve this information. Figure 15 shows a simple diagram that displays the relationship between products and ingredients.

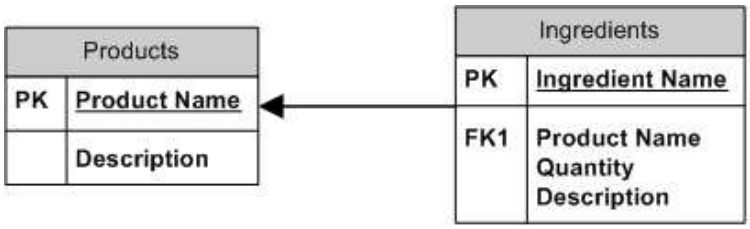

Figure 15 - Agrofood products ingredients

- **Supply chain and certification information:**

The supply chain and certification information are retrieved exclusively from the supply chain. Even though certifications are provided by third party organizations, producers usually hold to these information and are very eager to broadcast them, since researches have shown that costumers are very aware of these matters and give great importance to this kind of credential in the buying process. Figure 16 displays a diagram of products, their characteristics regarding the supply chain and their certifications.

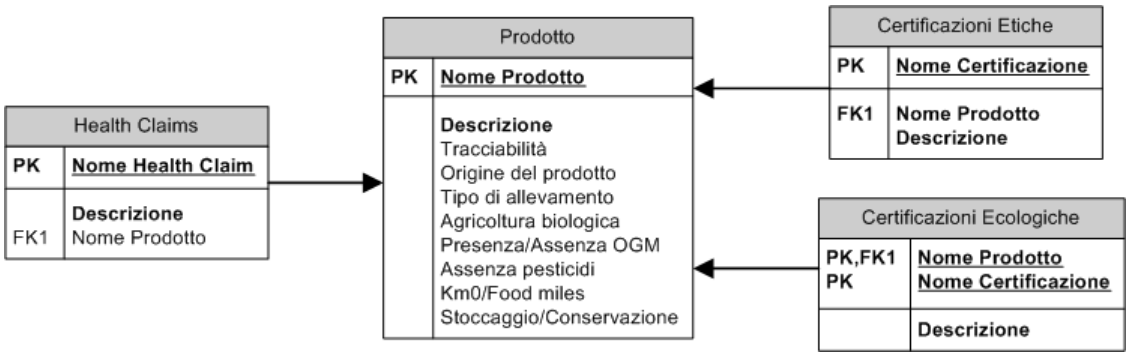

Figure 16 - Agrofood products and certificates

- **Recipes and ratings:**

Recipes and rating information are retrieved exclusively from the web. A good example of web sources with this kind of features is the American website FoodieView (FoodieView, 2011). This website allows users to search over one million recipes by ingredient, famous chefs, special diet considerations and type of cuisine. Also as a very interesting feature, recipes are rated and commented by users as to better evaluate the information provided. Pictures and videos are also available in order to instruct the users. It's always important to mention that all these features are offered for free. Figure 17 shows a simple diagram that explains the basic relationships between these entities.

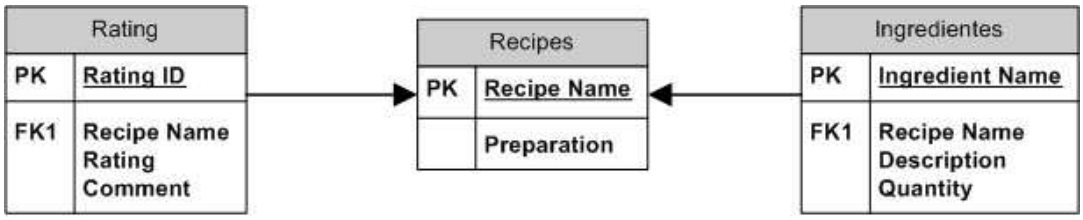

Figure 17 - Recipes and ratings

- **Restaurants and ratings:**

Restaurants and rating information are retrieved exclusively from the web. Once more, a very good example of web source for this kind of information is the American website Zagat (Zagat Survey, 2011). Provider of user-generated content, provides trusted and accurate restaurant ratings and curated restaurant reviews for thousands of top restaurants worldwide. Their robust restaurant search and rich free features help diners easily find the best restaurant for every occasion, anywhere, anytime. Restaurants are categorized by many features as to collaborate with better information filtering. Figure 18 shows a simple diagram that explains the relationships between these entities.

Figure 18 - Restaurants and ratings

Once we have wrapped the local sources we judge relevant for the system, we enter the mapping phase. As explained before, this step corresponds to the definition of a global schema and mapping of local schemas. This information is stored by linking local entities and their corresponding global entity. To understand a little bit better what is being said here, let's take a closer look at a couple of examples of source mapping applied to the Ms4A scope, based on the wrapped local schemas defined above.

Before we analyze the examples, It's important to keep in mind that in order to link local databases we must first identify pair wise common overlapping tables (borderline nodes) among the local schemas. Figure 19 generically represents this situation where S_1, S_2, S_3 represent the local schemas and n_j the borderline nodes overlapped.

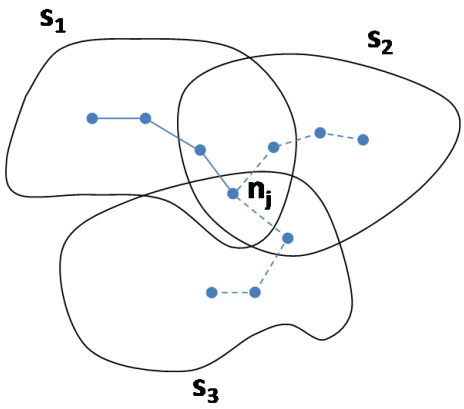

Figure 19 - Source mapping representation

- **Recipes / Dishes frontier tables:**

Analyzing the recipe and restaurant schemas we find out two possible overlapping tables, namely recipes and dishes. This first example already points out one of the main issues for global schema generation, i.e. the semantical similarities between entities with different names from different local data sources. In this case, in order to combine these two entities we could say that the dishes entity inherits the recipe entity information, since the last one basically describes the composition of the restaurant dishes and the process involved in making those plates. Even though both tables don't present exactly the same attributes, they have complementary properties and most important of all, they represent the same entity. One important thing to be noticed is that for attributes that exist in both entities, consistency needs to be checked in order to properly consolidate the data. By simply looking at Figure 20 it's easy to notice that the union of these two tables generates a more complex schema but at the same time, in return, generates a more complete, agile and easy to operate partially integrated database.

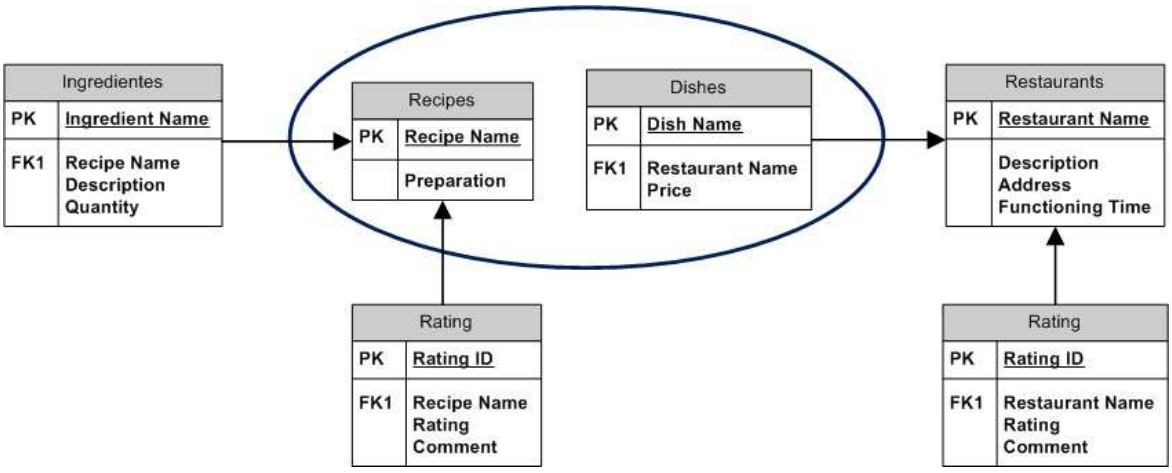

Figure 20 - Recipes / Dishes frontier tables

- **Product frontier tables:**

Analogous to the previous case, observing the nutritional, the ingredient and the supply chain schemas it's almost immediate to realize that these schemas can be easily linked together, since on each one of them we conveniently have a table named product that obviously describes exactly the same entity. As said before, in case of attribute replication along the integrated local data sources, data cleaning techniques must be applied in order to guaranty the quality of the information in the integrated database. This process can be really time consuming depending on the table sizes and especially on the number of local schemas being integrated. Figure 21 displays the unified diagram of these tables that as said before are not perfectly coincident but complementary, thus all the information is combined in the global schema.

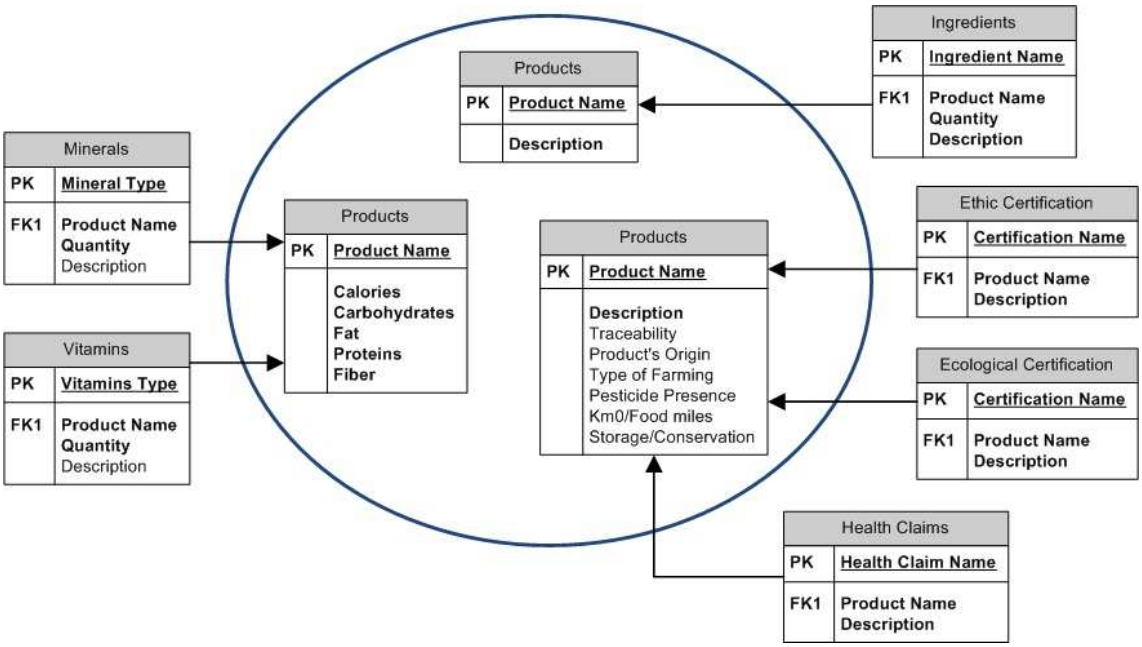

Figure 21 - Products frontier tables

- **Product / Ingredients frontier table**

Finally, we can still put together the two previously integrated diagrams to build a final integrated database that links all the information that we had wrapped before. It's important to notice that even though we suppose that the products studied can be composed by different ingredients; we consider the whole product as the ingredient for the recipes. For this reason, it's intuitive to understand that the ingredient table in the recipes schema can still be related to the products we are analyzing. Once again we face ourselves with semantic problems among data sources. Since the local data sources integrated in the Ms4A project are built for different purposes, not surprisingly the nomenclature used among them differs as to

adapt themselves to their primary purpose. In order to integrate schemas this semantic issues must be really clear in the designers mind, or otherwise inconsistent data can be generated by inappropriate mapping. Figure 22 shows in a simplified manner what was explained above. The final diagram of the integrated database is omitted due to its exaggerated proportions.

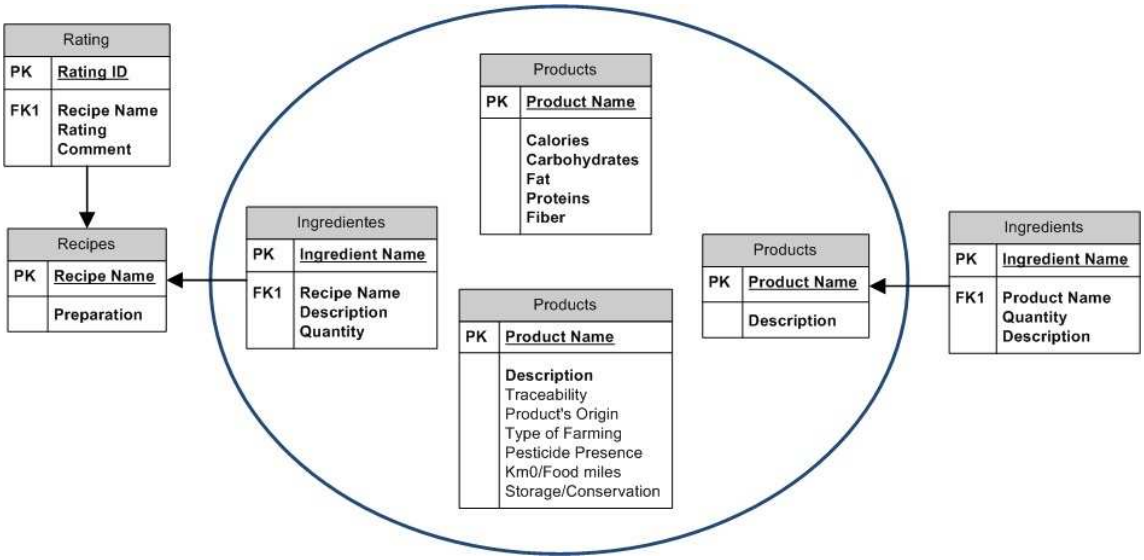

Figure 22 - Product / ingredients frontier tables

5.2. Model Analysis

Once the integrated database is built, what is left for us to do is test it and see if, as expected by the model, we can observe a considerable increase in information capacity, value and utility. To rudimentarly exemplify this idea, we propose a qualitative comparison between the typology of the queries usually applied to each local schema and the typology of the new application load of the integrated database.

Disconsidering data integration, it's easy to identify some classical examples of queries that can be executed by a system composed by the individual databases studied in the beginning of this chapter:

- How many calories, carbohydrates, proteins, etc. do I find in a certain portion of a product?
- Were pesticides used in the supply chain of this product?
- Is this product certified?
- What are the ingredients of this product?
- Is there any kind of allergen component in this product?
- What kind of recipes use this kind of product?

- Where can I eat a certain type of food (dish)?

Although the queries presented above are already really interesting and useful in a costumer point of view, data integration can increase all-around performance, utility, value and capacity of the queries executed. In the way we constructed our new database none of the properties of the previous local schemas are lost, and therefore, everything that could be done prior to the integration can still be done in the new system. On the other hand, all the effort we put into condensing all these data sources has to pay off somehow, and that is indeed what happens. Let's observe a few examples of the new enhanced queries that we are able to retreive from the new integrated database.

- Indicate restaurants that serve dishes with a certain product
- Indicate recipes that don't use a certain ingredient that might cause allergies.
- Display nutritional facts of a recipe or a dish in a certain restaurant.
- Suggestions of what to order in a restaurant based on recipes ratings.
- Indicate where can someone find restaurants that serve food made with certified products.

As we can see, data integration offers a considerable increase in sophistication of the queries that can be executed by the system, as well as an increase in the relevance of the information retrieved. Obviously, the greater the number of data sources linked the greater the options are.

It goes without saying that these new enhanced queries that we are able to retrieve from the new integrated database represent the increase in capacity that our system experienced. In order to better quantify how relevant this increase was, we would probably need some experimental data on the most common queries executed over a certain period of time (Application load). By classifying these queries into "old" and "new" queries, and comparing their absolute numbers, we can have an idea of the capacity gain that the data integration implementation generated. Since we don't have access to this kind of information, for now we just say that information capacity clearly increases in the integrated system.

Analyzing information utility is a really difficult task in this situation. Let's assume for simplicity reasons that if a query is requested by a user, the query has a positive utility effect. Analogously to what we did in the capacity analysis, in order to quantify the utility increase of the integrated solution we must compare the absolute numbers of "old" and "new" queries over a certain period of time. Since again we don't have access to this kind of information, for now we just say that information utility also clearly increases in the integrated system.

5.3. Cost Estimation

Last but not least, it's important to understand that even though there are great benefits associated to data integration, this processes also have costs, and in order to quantify these costs we subdivided them into parts so that they would be easier to understand and analyse. Doing some literature research and basing ourselfs on some practical cases we estimate that the costs involved in the implementation of a project such as Ms4A would be of the following typology and order:

1) Design time costs:
   - C_{wrap}: 750 € per source.
   - C_{map}: 100 € per common node consolidated between two different sources.
   - C_{maint}: 20% of the project's total cost.
   - C_{DQ}: Cleaning Tool (0 € - 800 €).
2) Run time costs:
   - C_{exe}: 10^{-4} € per query user.

Obviously, beside these operational costs that incur from the design operations in the integration process, there are also significant fixed costs related to hardware and software licensing that must not be disconsidered.

An interesting research recently conducted by Bloor Research (Howard P. , 2010) on Comparative costs and uses of Data Integration Platforms deals exactly with the analysis of this fixed costs that incur from the necessary information technologies acquisition in order to implement the desired integrated solution. A very extensive comparative analysis was made regarding the cost effectiveness of the main data integration solution providers, namely IBM, Informatica, Microsoft, Oracle, Pervasive, Hand Coding and Open source.

With the intention of showing how the costs of integration can vary considerably among different integration solution providers, the most relevant results retrieved from this benchmark study are displayed in the following graphs and tables.

Table 7 – Integration Project Initial Costs

	License costs	Additional hardware	Additional software	Implementation	Total first year costs
IBM	$175,781	$270,156	$51,094	$337,500	$834,531
Informatica	$ 41,089	$ 92,450	$32,507	$ 88,679	$254,726
Microsoft	$ 36,828	$ 37,591	$67,349	$ 62,599	$204,367
Oracle	$ 85,472	$104,018	$20,738	$132,578	$342,806
Pervasive	$ 29,390	$ 18,455	$21,873	$ 12,531	$ 82,250
Open Source*					$105,000
Hand coding*					$ 87,000

Table 8 - Integration Project Annual ongoing Costs

	Maintenance fees	Hardware	Admin	Internal tech staff	External consultants	Total annual costs
IBM	$75,750	$45,875	$74,438	$107,375	$75,063	$378,501
Informatica	$46,895	$31,368	$57,053	$ 86,000	$53,105	$274,421
Microsoft	$16,343	$21,969	$33,438	$ 48,156	$28,938	$148,844
Oracle	$30,500	$26,778	$35,222	$ 46,111	$42,778	$181,389
Pervasive	$23,194	$12,278	$18,944	$ 50,278	$11,056	$115,750
Open source*						$ 32,677
Hand coding*						$127,300

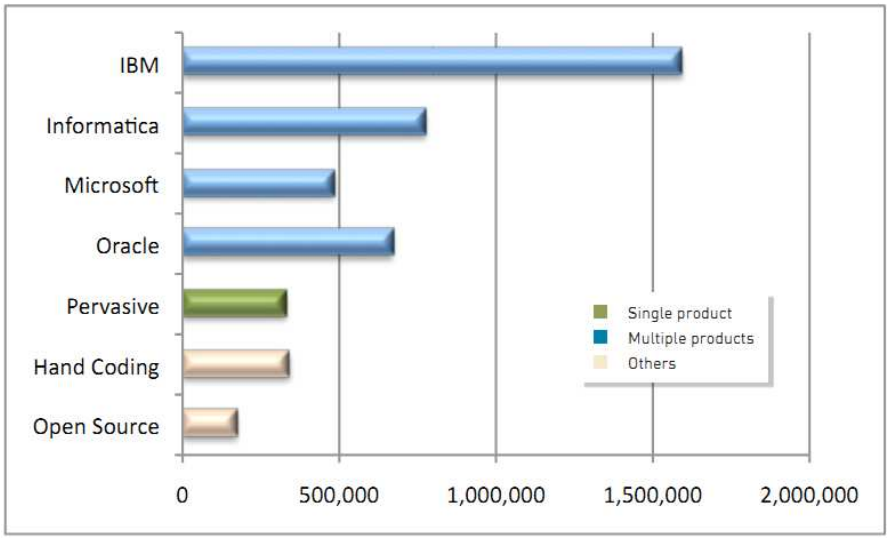

Figure 23 - Three-year Total Cost of Ownership (TCO)

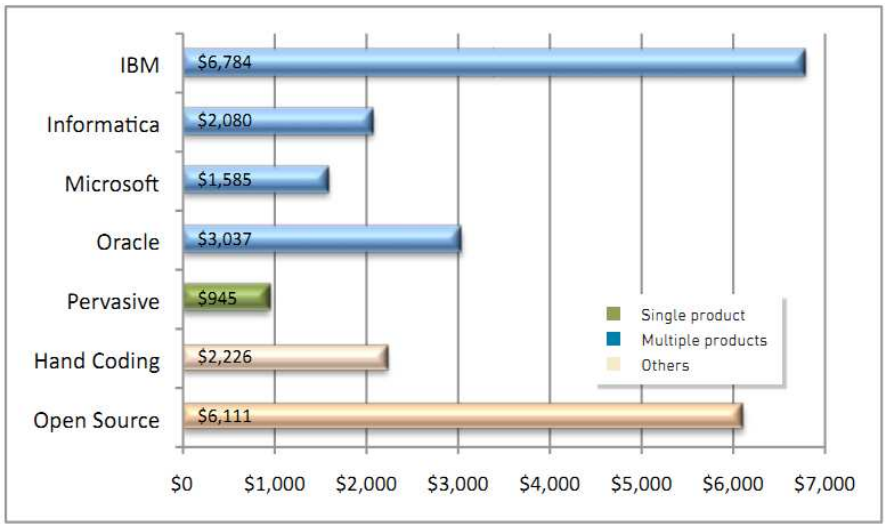

Figure 24 - Average costs (3-year TCO) per project per end point (sources and targets)

By looking at this somewhat intuitive statistics we realize the importance of choosing the right technology for the implementation of our project. Not all providers are the same and not always higher costs guaranty superior quality and performance. When implementing a data integration project we have to keep in mind that the technology to be used must be the one that best adapts to our context and that better fulfills our needs.

One final remark that needs to be made before we advance to the next chapter is that the exposed benchmark analysis considers fairly big integration endeavors in its statistics. For our application purposes we should probably stick to the statistics displayed on Figure 24. For some additional information on this cost analysis study please refer to the appendices section of this paper.

6. DISCUSSION

In this chapter we take the time to point out some of the problems and issues that we ran into during the design and analysis of the presented integration solution.

As shown in the previous chapter, wrapping and mapping are pretty straight forward processes. Even though there are several approaches for doing these things, as long as all the operations are carefully executed, the results in this phase shouldn't change much, especially when the number of local sources being integrated is not that big.

It's important to realize that one of the biggest problems we faced in this project was the lack of concrete data on the Ms4A scenarios. The absence of this information made it significantly difficult to quantify the variation of the parameters analyzed by the economic model, namely, value, quality, utility and capacity. Even though we are able to identify an increase or decrease in these parameters, it's really hard to assess the overall changes that each one suffered after the integration process.

According to the model presented, in order to evaluate the economic value of an integrated solution, it's imperative to compare the project's utility and costs variables. Because of its subjectivity, it's really hard to monetize the utility of a given project, and therefore, the comparison suggested in the model is really complicated and can only be done in a qualitative manner. Although we cannot quantify the information value acquired by the data integration solution, that doesn't mean that we cannot determine if it has grown or decreased, and that fact is already of great interest.

What we notice at this point is that even though the model is pretty consistent and logically built, it doesn't ponder on the subjectivity of the variable considered to evaluate the information.

For all the reasons mentioned above, our study takes a more qualitative and empirical approach of the analyzed problem. We are not able to present precise figures on the utility, value, capacity and costs of this project but we can identify trends on the analyzed variables and this is enough for an educated and reasonably accurate conclusion on the feasibility of this project.

7. CONCLUSION

Organizations tend to create databases of interest through a series of projects and realizations that result in a database architecture characterized by a set of anomalous behaviors. Such behaviors concern the redundancy of representations, the misalignment of data among different databases, the in-coherence in business rules related to the same objects in different databases, and errors in data that result in the heterogeneous representations of records pertaining to the same real world object. This trend is made more and more critical by the continuous evolution of organizations due to merger and acquisition activities. Consequently the problem of managing the whole data architecture by migrating from traditional DBMS technologies to integration technologies is a primary issue in modern organizations.

In this study, despite all the theoretical grounds that were provided on data integration concepts and economic models for data integrated architectures, we were able to accomplish three main things (as established in our goals).

Firstly, we were able to design/model the data integration solution for the Ms4A project based on the presented reference scenarios. Through this process we analyzed the main issues involved in the wrapping of local data sources and mapping of a global schema. Although always highlighted, data quality issues were left aside because of unexisting real world data in the modeled local data sources.

Secondly, we applied the quality based model for information value and capacity in the integrated database modeled in step one. Since quantifying abstract concepts such as information quality, utility, capacity and value is an extremely hard and imprecise activity, we opted for a more empirical and qualitative analysis of these variables. By demonstrating the increase in the number of the queryable scenarios in the global schema, we proved that integration initiatives indeed increase the information value, assuming that all data quality procedures were properly applied in order to guaranty minimal error propagation in the integrated schema. The utility variable was briefly covered in the parallel study conducted by DEPAAA in the identification of the reference scenarios (see Chapter 2), where we concluded that in the eyes of the users, the integration initiative indeed collaborated for an increase in the information utility.

Thirdly, we benchmarked development costs for data integration solutions. The cost analysis was divided in “modeling and design” costs and “hardware and software licensing” costs. The first ones have a more variable cost characteristic since the costs here are usually correlated to the number of local data sources that are being integrated. The second one has a

more fixed cost characteristic since the cost here usually reflect the technological choices made in order to enable the projects execution.

Based on all the presented results we can clearly say that the Mobile Service for Agrofood project is indeed a viable project that according to the analysis conducted in this study has actual marketable economic value. Therefore it's safe to say that implementing this solution under the standards here presented is a considerably safe and smart investment.

Issues related to who should carry on the development of this application are still an ongoing discussion. It's obvious that clients, producers, supermarkets, etc. all benefit from the implementation of this project. Researches have shown that costumers are actually resistant to possible increases in the products price in exchange for information. Current scenario seems to indicate that if any investment is made for the implementation of this project it should probably come from producers or big chain supermarkets.

8. RECOMENDATIONS

Data integration is such a rich field that several important aspects not addressed in this research can be identified and applied to the Ms4A scenarios in order to retrieve more valuable information for evaluating critical aspects of this project.

A simple extension point for this research could be a prototype study of the designed integrated database architecture exposed in this thesis. By actually trying a simplified hands-on approach of data integration in the Ms4A scenarios, new interesting issues might come up, as well as more technical factors could be analyzed.

Another possible future topic of study regarding the Ms4A project could be an analysis of the benefits and issues of coupling integrated database with data mining techniques to yield even more interesting and useful information. This information could be used either by users or by producers (for further focus of their targeted marketing campaigns).

Finally, something that also draws some attention is the possibility of extending this project's service capabilities towards an advisory system, i.e. a system that based on the users profile, the data available in the integrated database and a historical record of the users' queries, would be capable of suggesting alternative products (or restaurants, or recipes) to the user, based on a specific algorithm. Since most of the infrastructure is already available, seems reasonable that this possibility should be also studied.

REFERENCES

Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applies Systems Analysis* .

Adelman, S., & Moss, L. (2003). What are the advantages and disadvantages of corporate data integration? *Information Management Online* .

Bantere, A., & Cavaliere, A. (2011). *An economic model of MOSE services in the food industry.*

Batini, C., & Lenzerini, M. (1986). *A Comparative Analysis of Methodologies for Database Schema.* Rome: Dipartimento di Informatica e Sistemistica.

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques.* Springer.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). *Methodologies for data quality assessment and improvement.* ACM Comput. Surv.

Bellinger, G., Castro, D., & Mills, A. (2004). Data, Information, Knowledge, and Wisdom. *A journey in the realm of systems* .

Black, P. E. (2004). *Dictionary of Algorithms and Data Structures.* U.S. National Institute of Standards and Technology.

CalorieKing Wellness Solutions, I. (2011). *Calorie King Food Search*. Acesso em 20 de 11 de 2011, disponível em Calorie King: http://www.calorieking.com/

Cappiello, C., & Francalanci, C. (2011). Information value and information capacity: a unified quality based model. *International Conference on Information Systems 2011.*

Dittrich, K. R., & Jonscher, D. (1999). *All Together Now — Towards Integrating the World's Information Systems.*

Domenig, R., & Dittrich, K. R. (2000). *An Overview and Classification of Mediated Query Systems.*

Fenner, A. (2002). Placing Value on Information. *Library Philosophy and Practice* .

FoodieView, T. (2011). *FoodieView Recipe Search*. Acesso em 20 de 11 de 2011, disponível em FoodieView: http://www.foodieview.com/

Gantz, J. F. (2008). *The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011.* IDC - EMC.

Gartner. (2008). Key Cost-Cutting Tactics in Data Management and Integration. *Gartner Business Intelligence Summit* .

Halevy, A. Y. (2005). Enterprise information integration: successes, challenges and controversies. *SIGMOD* .

Howard, P. (2010). *Comparative costs and uses of Data Integration Platforms.* Bloor Research.

Howard, R. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics* .

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling .*

Lenzerini, M. (2002). *Data Integration: A Theoretical Perspective.*

Levy, A. S., & Fein, S. B. (1998). Consumers' Ability to Perform Tasks Using Nutrition Labels. *Journal of Nutrition Education* .

Moody, D., & Walsh, P. (1999). *Measuring the value of Information: An asset valuation approach.* ECIS.

Moss, L., & Adelman, S. (2000). *Data Warehouse Project Management.*

Nayga, R. M. (1996). Determinants of Costumers' use of nutritional information on food packages. *Journal of Agricultural and Applied Economics* .

OECD. (2004). Organization for Economic Co-operation and Development. *Heath Data* , Paris.

O'Reilly, T. (2005). What is Web 2.0. *O'Reilly Network* .

Tsierkezos, S. (2010). *Comparing Data Integration Alorithims.* Manchester.

Wang, R., & Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* .

Wiederhold, G. (1992). *Mediators in the architecture of future information systems.* IEEE Computer.

Zagat Survey, L. (2011). *Zagat Lists*. Acesso em 20 de 11 de 2011, disponível em Zagat: http://www.zagat.com/

Ziegler, P., & Dittrich, K. R. (2005). *Three Decades of Data Integration - All problems solved?*

APENDICES

A. Comparative costs and uses of Data Integration Platforms

The integration scenarios analyzed were basically 6:

- Data conversion and migration projects
- ELT, CDI, and MDM solutions
- Synchronization of data between in-house applications (such as CRM/ERP)
- Synchronization of data with SaaS applications
- B2B data exchange for costumer/supplier data
- Implementation of SOA initiatives

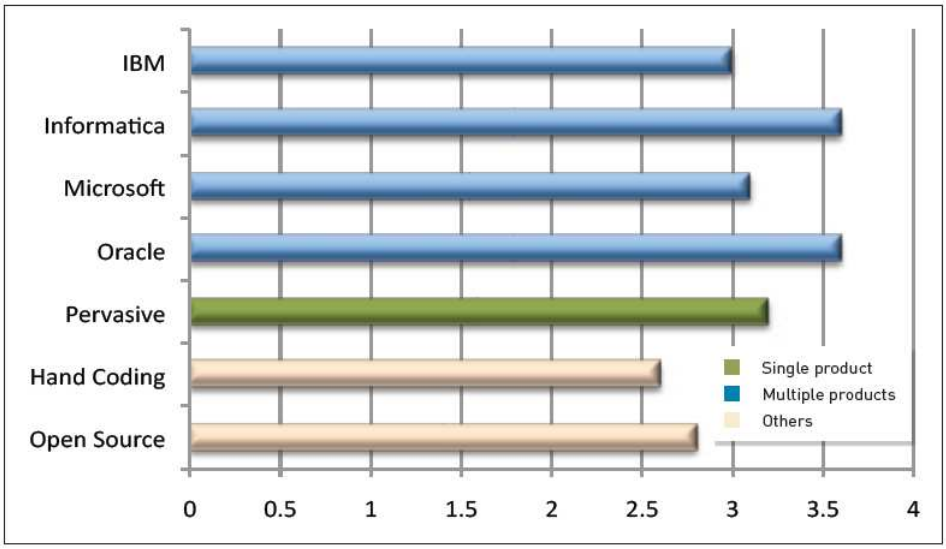

Figure 25 - Average number of scenarios for which products/vendors are considered suitable

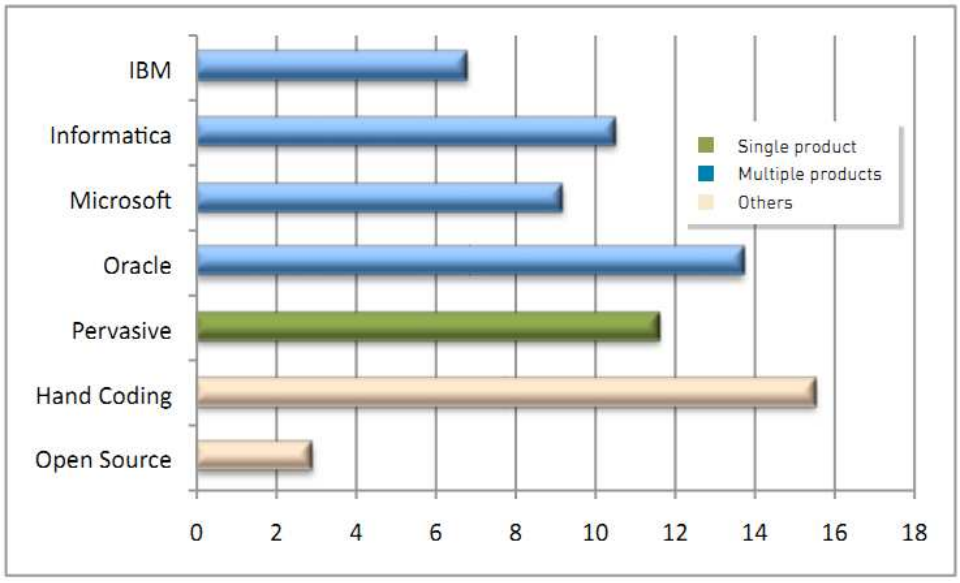

Figure 26 - Average number of end points (sources and targets) per project

Table 9 - Ramp up time and effort

	Time to learn	Resources required to build first solution		
	Weeks	Internal staff (man weeks)	External consultants (man weeks)	Total (man weeks)
IBM	7.3	10.0	6.8	16.8
Informatica	4.2	7.2	5.1	12.3
Microsoft	4.3	7.4	3.0	10.4
Oracle	6.5	11.9	5.1	17.0
Pervasive	3.0	5.6	2.5	8.1
Hand coding	4.6			5.6
Open Source	6.5			9.8