On semi-automated matching and integration of database schemas

Ünal Karakaş, Ö.

Publication date
2010

Chapter 1

Introduction

To effectively use and benefit from the vast amount of information provided online by a large number of databases, this information needs to be interlinked and integrated. This requirement has become more evident with the increasing demand for remote collaboration among independent organizations and individuals. An important first step in this direction is to support the matching of independently developed meta-data in database schemas of different organizations and individuals. This needs to be done through resolving variety of their heterogeneities, and identifying correspondences among concepts defined in these database schemas. Furthermore, for proper interlinking of these databases, another necessary step is the integration of their database schemas. Resolving these complexities is challenging and quite inefficient to handle manually. The thesis proposes an automated but supervised approach, called SASMINT- Semi-Automatic Schema Matching and INTegration, which addresses and merges the problems of matching and integration of relational database schemas. This chapter provides some introductory information about the research work carried out towards provision of the proposed approach. Section 1.1 addresses the motivation for this research. Section 1.2 enumerates the main research questions, followed by the main objectives and contributions of the research addressed in Section 1.3. Section 1.4 specifies the scope of this research. Finally, Section 1.5 elucidates the applied research method, and Section 1.6 outlines the structure of the thesis.

1.1 Motivation and Requirements Analysis

Advances in information and communication technology (ICT) have created new opportunities for computing world-wide. High speed networks enable us to reach large quantities of information within fraction of seconds. However, these developments create many new challenges. One such example is how to link and share large amounts of similar or inter-related data provided by distributed, heterogeneous, and autonomous parties who wish to work with each other.

The importance of developing a support infrastructure for data sharing has been addressed and understood clearly during the last decade, with the increasing need for collaboration among organizations. The term collaboration among organizations is now used frequently, as defined in (Camarinha-Matos & Afsarmanesh, 2008b):

> **Collaboration** is a process in which entities share information, resources and responsibilities to jointly plan, implement, and evaluate a program of activities to achieve common goals.

For companies for instance, in order to remain competitive in a highly aggressive global market, they need to become more agile in coping with changes and achieve this goal in a better and faster manner. As a response to this challenge, Collaborative Networks have emerged.

> A **collaborative network** (CN) is an alliance constituted by a variety of entities (e.g. organizations and people) that are largely autonomous, geographically distributed, and heterogeneous in terms of their operating environment, culture, social capital and goals, but that collaborate to better achieve common or compatible goals, and whose interactions are supported by computer networks (Camarinha-Matos & Afsarmanesh, 2008a; Camarinha-Matos et al., 2005).

Several forms of collaborative networks are currently observed, as shown in Figure 1.1. The two top level collaboration forms in this classification are the ad-hoc collaboration and Collaborative Networked Organizations (CNOs). The ad-hoc collaboration represents formation of spontaneous collaborations, without any predefined goal, such as the instantaneous on the spot formation of a rescue collaboration team to assist with a disaster. On the other hand, CNOs are carefully established with participating organizations, having different roles in the network, towards achieving their common goals. There are also two

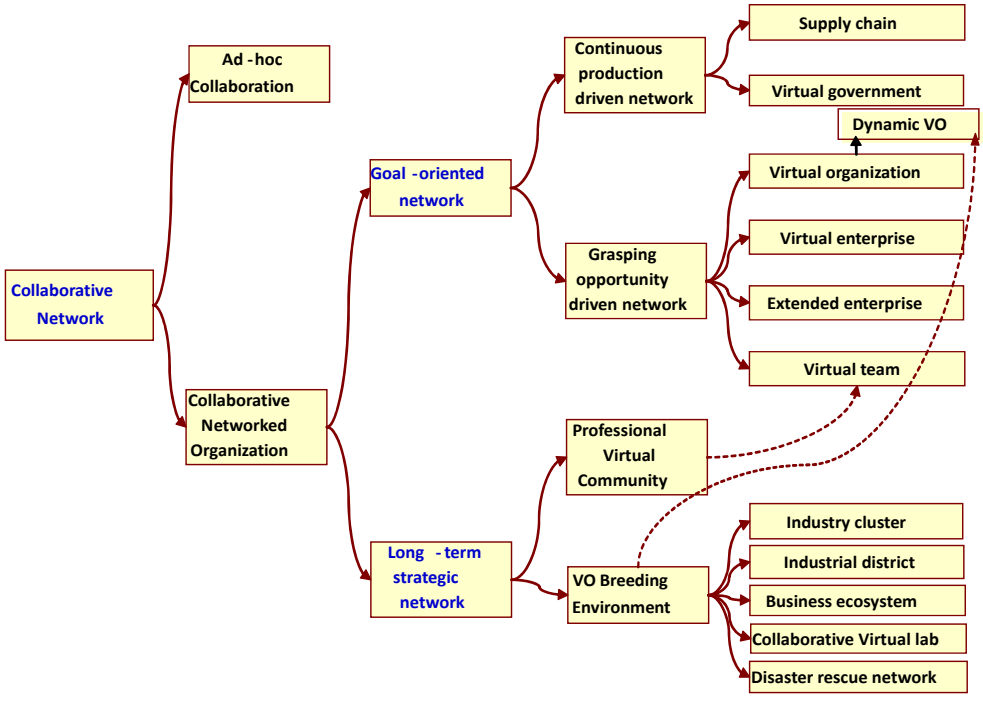

Fig. 1.1. Different forms of collaborative networks (Camarinha-Matos & Afsarmanesh, 2008a)

forms of CNOs called goal-oriented and long-term strategic networks. Members of a goal oriented network work together to achieve their common goals, whereas long-term strategic networks are strategic alliances aimed to prepare their member organizations towards dynamic establishing of focused collaborative networks at the emergence of new opportunities in the market/society. Goal-oriented networks can be either continuous production driven networks or grasping-opportunity driven networks, while long-term strategic networks can be either professional virtual community or virtual organization breeding environment (Afsarmanesh & Camarinha-Matos, 2005). At the lowest level of classification, supply chain, virtual government, virtual organization, virtual enterprise, extended enterprise, virtual team, industry cluster, industrial district, business ecosystem, collaborative virtual laboratory, and disaster rescue network represent the main variety of types of networks that are manifested today in parallel. Details about each of these types of collaborative networks are provided in (Camarinha-Matos & Afsarmanesh, 2008a).

We take Virtual Organizations as an example. Briefly, a *Virtual Organization* (VO) is a gathering of autonomous organizations through a network that pursue the accomplishment of a set of specific common goals (Camarinha-Matos et al., 2000). There are a number of benefits associated with VOs (Afsarmanesh & Camarinha-Matos, 1997), including the increased access to market/society opportunities, sharing risks, reducing costs, and achieving business/societal goals not achievable by a single organization and thus the motivation for involvement in VOs. A VO represents a complex and dynamic entity that undergoes a sequence of stages during its life cycle (Camarinha-Matos & Afsarmanesh, 1999a; Camarinha-Matos & Afsarmanesh, 1999b), as shown in Figure 1.2.

In all stages of the VO lifecycle, ICT support is needed. For example, in order to enable rapid formation of VOs, it is required to establish a common interoperable infrastructure (Afsarmanesh & Camarinha-Matos, 2005) that is typically achieved within the VO breeding environments (Afsarmanesh et al., 2008). Members of VOs need to strongly interact with each other to achieve the goals of the VOs and one form of this interaction is by means of data

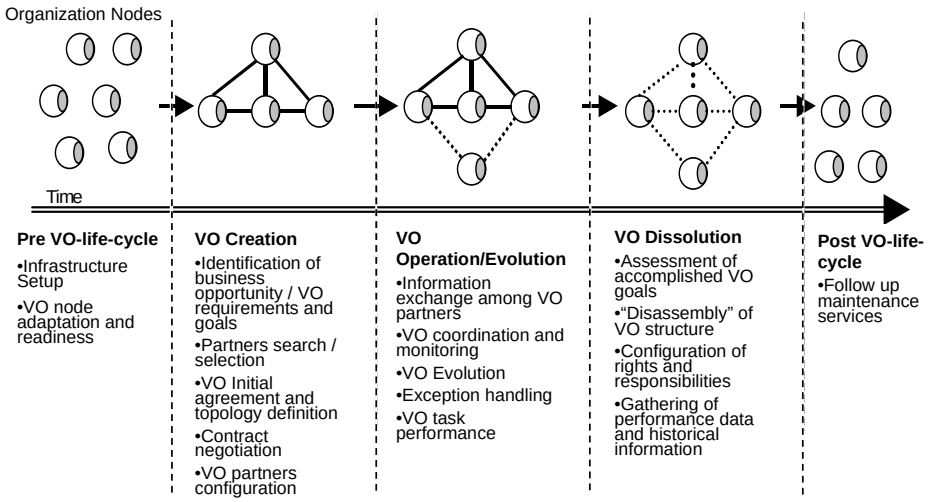

Fig. 1.2. VO Life Cycle

sharing, which requires information integration and interlinking.

Although in practice the first examples of collaborative networks come from the manufacturing domain, the need for collaboration has been well understood recently in different domains, such as engineering, economy, social sciences, etc. One such example domain is the biodiversity. Increasing number of biodiversity conservation activities entail producing more accurate results by comparing and/or merging different biodiversity data analysis results to make better predictions about the global status and distribution of species, as well as protection of those who are endangered.

This, in turn requires the collaboration and data resource sharing among many worldwide distributed biodiversity centers, organizations, and individual researchers (Unal & Afsarmanesh, 2006a; Unal & Afsarmanesh, 2006b; Unal & Afsarmanesh, 2006c; Unal & Afsarmanesh, 2009; Unal & Afsarmanesh, 2010). Although the importance of collaboration in biodiversity has become clear to most involved scientists, so far most biodiversity related organizations hesitate to actively cooperate. This is mostly due to the sensitivity of some specific data categories, such as those related to endangered species in biodiversity domain, where there is the danger of unintentionally creating new business opportunity for illegal poachers, though announcing the information about species in danger of extinction. Therefore, new mechanisms and infrastructures are needed, supporting information sharing among organizations, while taking the needed criteria into account. With the existence of such mechanisms, organizations can more easily decide to collaborate.

There are however some difficulties facing the infrastructures to support data sharing and exchange in the biodiversity domain as follow. Different organizations structure biodiversity data in different formats depending on their specific needs and preferences. The level of detail that they keep for their managed data also greatly differs. They typically do not use standard data models and different syntax is used. Likewise, since different centers use their own controlled terminology and vocabularies, the semantic definitions of their data and used concepts are heterogeneous. These matters have resulted in a large number of independent and heterogeneous databases, scattered all over the world. Because of the differences in managing their data, these databases are quite rarely integrated effectively. Furthermore, any effort spent on such integration is usually not at the global network level, rather bi-lateral focused on each pair of databases. Hence, demands for an effective uniform mechanism to integrate/interlink a number (possibly large number) of databases, to support homogeneous access to heterogeneous and distributed databases, thus providing a single and integrated interface for users in biodiversity networks are increasing.

As discussed above, in biodiversity domain, collaborating organizations need to share data with others and simultaneously access and manipulate data from others, and thus integration of data is required. Among other challenges, the heterogeneity, distribution, autonomy, continuous and rapid technologic evolution, and multi-disciplinarity of the area are the main obstacles faced to achieve the required integration in all levels of collaborative networks (Camarinha-Matos & Afsarmanesh, 2003), including those of the biodiversity. We can further elaborate on these common obstacles as follows: 1) *Heterogeneity*: It arises due to the lack of common standards, as each organization uses its own format and model for database schema definitions, which makes the interoperation among nodes much more difficult. Especially if the number of organizations in a CN is large, an important challenge is how to share and integrate data represented by heterogeneous database schemas. 2) *Distribution*: Organizations are logically and physically distributed. However, improvements in the field of high-speed networking help to decrease the impact of this obstacle and properly support remote access to distributed databases. 3) *Autonomy*: Organizations autonomously decide what to share and with which other organizations. Furthermore, each data owner in CNs is autonomously deciding on the representation and modeling of its data, which clearly and vastly increases the

heterogeneity. 4) *Continuous and Rapid Technologic Evolution*: Fast improvements in technologies lead to continuous changes in the format and amount of data to be exchanged. This evolution further increases the heterogeneity problems. 5) *Multi-disciplinarity*: Since each organization is from different types of areas and disciplines, integration and interlinking of information needs to handle wide variety of information types and their specificities. Design and implementation of an integration infrastructure for CNs, needs to take all these obstacles into account.

Heterogeneity is the most relevant obstacle among the others facing the data sharing for collaborations. The collaboration infrastructure has to consider such differences for providing effective mechanisms to integrate/ interlink and for homogeneous access to databases. Rather than accessing and manipulating single database systems in isolation for CNs, database research is needed to address simultaneous access and manipulation of different remote databases, as suggested in federated databases and multidatabase approaches (Hammer & Mcleod, 1979; Heimbigner & Mcleod, 1985; Sheth & Larson, 1990). However, automatic resolution of schema heterogeneity still remains as a major bottleneck for provision of integrated data access/sharing among autonomous, heterogeneous, and distributed databases.

Since each data owner in CNs decides autonomously on the representation and modeling of his data, data are typically and widely heterogeneous even if from the same domain. In order to provide transparent access to such remote data and enable the sharing of information among databases, their schema heterogeneity needs to be identified and resolved and then the correspondences among different organizations' schemas need to be identified. This process is called in research as schema matching. After schema matching process, to support collaboration, e.g. the possibility of processing federated queries within the network, schemas usually need to be integrated, to facilitate the needs of the CNs. It is clear that schema matching and integration constitute two key processes of the ICT infrastructure supporting the collaboration among organizations. Thus, tools that enable semi-automatic matching and integration are among the most important components of such infrastructures.

The most difficult part in resolution of heterogeneity of schemas during the schema matching process is the identification of the semantics introduced at each organization that are incorporated into their schema definitions. Data semantics are related to database designers' preference or interpretation of data, according to their understanding of the world within their organizations. Different interpretations cause different representations of data and thus different data models. In the process of comparing different database schemas for instance, semantic heterogeneity arises out of the ambiguity inherent in the separation between names (words) chosen in a data model and what they represent within the organization that originates them. Some semantics in general can be inferred from data, schema, and annotations if they exist. However, to put it simply, this information is most of the time incomplete and such inferences are not fully accurate, to decide whether an element *x* of a schema A matches an element *y* of another schema B and there is no other element *z* of this second schema B that matches *x* better than *y*. Therefore, a fully automatic solution for schema matching may not produce the best results.

Approaches so far proposed in database research for providing access to distributed, heterogeneous, and autonomous data sources have addressed some aspects of semi-automatic schema matching and/or schema integration in their approach. However, these approaches suffer from some or all of the following main limitations:

- No approach deals with both schema matching and schema integration together. Furthermore, it is generally not addressed how to formalize the result of schema matching and how to facilitate and support the needed semi-automatic schema integration.

- A fully automatic schema matching and integration is not realistic, considering that some types of semantic and structural conflicts are difficult to resolve automatically, as addressed later in Chapter 3. Therefore, a simple but effective user interface is needed to enable user interaction with the system for modification of the results generated automatically for both schema matching and schema integration. Nevertheless, in the proposed approaches so far, the provision of needed user-friendly interface as a part of the proposed architecture is typically skipped.
- As for schema matching, the suggested approaches typically represent at least one of the following drawbacks:
    - Although the aim is to automate or semi-automate the matching process, the currently proposed solutions generally require too much manual work.
    - A limited number of algorithms are so far implemented each focused on the automatic resolution of certain specific challenges related to either syntactic, semantic, or structural conflicts, and there are no comprehensive solutions suggested. As explained in Chapter 3 and addressed by most other work, syntactic, semantic, and structural conflicts constitute the main categories of heterogeneities that exist among database schemas. While observing both the nature of existing schemas, which generally consist of elements with different syntactic, semantic, or structural characteristics, and our test results presented later in Chapter 6, in each case, using a combination of some of these algorithms, which are suitable for different types of elements and domains, is necessary for achieving more accurate results.

There are a number of key requirements that an ICT infrastructure for CNs needs to address and support for enabling the data sharing and exchange among organizations. Namely, the base requirements listed below, must be met independent of any specific solution for data sharing (Guevara-Masis et al., 2004; Unal et al., 2005).

- Organizations should be able to preserve their autonomy when they join a collaborative network. They should be able to autonomously decide which part of their data and with which other nodes to share.
- New organizations should be able to join the networks easily, and dynamic evolution of the schemas, representing the shared data, should be supported.
- Database administrators should be supported with tools to semi-automatically generate mappings from each of the different schemas to the integrated schema.
- Organizations in CNs should easily collect data from others without needing to deal with the underlying heterogeneities of databases.

Two of the most important components of an ICT infrastructure, which meet the requirements above for CNs, are the processes and components for schema matching and schema integration. Namely, the local schemas of a number of organizations need to be semi-automatically matched and integrated to generate a global schema for the CN.

Schema matching and integration play important roles in providing data sharing among distributed, autonomous, and heterogeneous databases. Taking into account the limitations of existing approaches, a comprehensive solution for semi-automatic schema matching and integration needs to focus on a number of specific requirements, as addressed below:

- **Semantic information needs to be identified:** Inherent in the schemas are large amounts of semantic information. Identifying semantic relationships is harder than simple relationships, as there are more possibilities that need to be taken into account. While observing the explicit relationships among schema elements, identifying implicit relationships is a problem that makes the automatic detection of elements' correspondences difficult. Auxiliary resources, such as linguistic dictionaries consisting of some semantic relationships among concepts, need to be utilized to identify as much semantic information as possible.
- **Both simple and complex matches need to be considered:** Most matching approaches limit their search to only one-to-one (1-to-1) matches (e.g. "email" to "electronic_mail"), also called as simple matches. Complex matches (e.g. "address" to "street", "zipcode", and "city") are much more difficult to identify than 1-to-1 matches. Although it is not realistic to extract all variations of matches automatically, at least complex matches in form of 1-to-n and n-to-1 need to be also identified to the extent possible.
- **Combination of a number of matching algorithms needs to be considered:** Schemas in general consist of element names in different formats. Some similarity algorithms produce better results when applied to certain specific types of element names. Therefore, it is not effective to pre-select and use only one or a few comparison algorithms, which are each suitable for certain types of names, for all kinds of schemas.
- **A Supporting user friendly Graphical User Interface (GUI) needs to be provided:** Developing only algorithms for automatic schema matching alone is not sufficient. User interaction is an important part of the process to be considered when developing the schema matching and schema integration systems. Especially considering that it is not possible to identify all matches automatically, a user-friendly and effective user interface is required to enable users' modification of the matched results. Furthermore, the results of schema integration also need final users' validation.
- **Schema matching process needs to be combined with schema integration process:** Schema integration, a challenging process especially considering all the conflicts that need to be resolved before the integration starts, requires at its base the identification of the correspondences among the source and target schemas, resulted by the schema matching. Therefore, a schema integration approach should facilitate the schema matching process by formalizing its user validated results and applying these results to the schema integration process. Proposing such a semi-automated schema integration approach and implementing it as a system provides a significant contribution to the information sharing and integration within the CNs.

1.2 Addressed Research Questions

In Section 1.1.1, we classified the general data sharing requirements for CNs under the umbrella of schema matching/integration. Namely, we addressed the CN's related requirements to support data sharing among distributed, autonomous, and heterogeneous databases. Addressing this problem area, we aim at developing formally founded and empirically validated approach and mechanisms. As such the main General Research Question (GRQ) for this thesis constitutes:

> ***GRQ-*** *How can we effectively and semi-automatically achieve the schema matching and schema integration, to facilitate data sharing in Collaborative Networks?*

We further refine this general research question into four specific Research Questions (RQs), which are addressed by this thesis.

In the first question, we address the terminology used in database research related to approaches and architectures for data sharing among heterogeneous data sources. This leads to the required understanding of the domain of our research problem:

> ***RQ1-*** *Which effective approaches and architectures can enable data sharing through interlinking and/or integrating heterogeneous databases of distributed nodes?*

Heterogeneity is the main problem to be tackled when dealing with schema matching and integration. It is therefore, necessary to differentiate the potential types of heterogeneities in order to identify those on which we need to focus during the schema matching and integration processes. This leads to our second research question:

> ***RQ2-*** *What is a representative taxonomy for addressing database schema heterogeneities, and in turn applicable to formalization of schema matching and schema integration challenges?*

Based on the state of the art and currently open research issues, we need to propose an appropriate approach for enabling semi-automatic schema matching and integration. This approach therefore should semi-automatically resolve different kinds of schema conflicts, such as the syntactic, semantic, and structural conflicts, in order to identify the potential matches among schema elements. The approach should be verifiable, e.g. a proof of concept as a working prototype of the system needs to be developed. Another important point that needs to be supported is the design of proper 'User Interaction'. These points lead to our third research question:

> ***RQ3-*** *What are effective mechanisms for semi-automatic schema matching and schema integration, and how should the user be involved in the process?*

We think the validation is important and necessary, in order to indicate the 'accuracy' and effectiveness of the approach we propose in comparison to other work. So, the final research question is built around the challenge of validating the developed system. The answer to this question shall reveal the appropriate measures that can be used for evaluating the accuracy of schema matching and schema integration, and thus the fourth research question constitutes:

RQ4- *How can we assess and validate the effectiveness of the proposed semi-automatic approaches for schema matching and schema integration?*

1.3 Objectives and Contributions of the Thesis

Aiming to address the main general research question described in Section 1.2, the main objective of this thesis is to propose an approach for resolving syntactic, semantic, and structural conflicts for semi-automatic schema matching and integration, facilitating data sharing and exchange in CNs. The answers to four specific research questions form the objectives of this thesis. Namely, the first research question (RQ1) is addressed by analyzing the related architectures and terminology used in database research for data sharing among heterogeneous data sources, which is the main subject of Chapter 2. The second research question (RQ2) is addressed by analyzing different taxonomies of heterogeneities proposed in the literature and defining the taxonomy of heterogeneity related to the challenges for schema matching and integration. This is the subject of Chapter 3. Our approach for semi-automatic schema matching and integration and its implementation are described in Chapter 4 and Chapter 5 in order to meet the research question three (RQ3). Research question 4 (RQ4) is addressed by carrying out validation against other related research. Chapter 6 describes this validation work and its results.

To conceptually verify our approach, we design and implement the SASMINT system that forms the basis for an infrastructure enabling users to query heterogeneous and distributed databases transparently in a federated database environment. Based on the proposed approach and its implementation, the main contributions of this thesis can be listed as follows:

- ✓ **Supporting both simple and complex matches:** Unlike many other approaches that support only simple matches, SASMINT supports both simple and complex matches, as addressed before.
- ✓ **Elevating the accuracy of schema matching:** In the SASMINT approach and implementation, we utilize a weighted combination of several schema matching algorithms. Syntactic, semantic, and structural conflicts are resolved by applying different specific string and structural similarity algorithms rooted in Natural Language Processing (NLP) and the Graph Similarity domains. Each algorithm best suits a specific type of strings and graph structures, and thus compounding some of them in SASMINT gives rise to more accurate matching results than other proposed approaches.
- ✓ **Enabling semi-automatic schema integration:** SASMINT interrelates directly the schema matching results with the schema integration. Heuristic rules are defined that run on the results of the schema matching and generate derivation formalism for an integrated schema automatically. We assess this as a novel contribution providing a strong competitive edge for the research on the SASMINT system.
- ✓ **Definition and incorporation of an XML-based language (an XML Schema) for enabling unambiguous interpretation of schema match / integration results:** Within the SASMINT system, we have devised an XML-based derivation language, which we call the SASMINT Derivation Markup Language (SDML) (in the format of XML Schema), that captures and supports the creation of a persisting schema

match and schema integration results**.** The value proposition of this particular contribution is multi-faceted. First, the persisted schema match integration results enable the external systems/agents to unambiguously interpret/understand the match / integration results. These external systems/agents could consume this information for implementing federated query processing, etc. Second, this generic format is understandable by the match/integration related human agents in-the-loop. What this means is that the human agents can then easily modify these results. Finally, the structure of the derivation language is designed to keep the derivation history. Namely for every entity, its entire derivation tree is preserved. This feature in turn enables the incremental schema integration procedure.

- ✓ **Enabling semi-automatic identification of suited weights for the composed algorithms:** A number of algorithms are utilized in the composite approach of the SASMINT system that calculates their weighted sum. Therefore, it is important to assign an appropriate weight to them, bearing in mind the suitability of each algorithm for different types of inputs. SASMINT provides the SAMPLER technique to semi-automatically identify the appropriate weights for the algorithms used in the linguistic matching.
- ✓ **Enabling user-friendly interaction by means of a GUI editor:** It is not possible to automatically extract all types of semantic and resolve all kinds of structural conflicts. Therefore, a suitable user-friendly GUI editor is provided for SASMINT for supporting the visual modification of the results of both schema matching and schema integration processes as well as their storage for further use.

1.4 Scope of the Research

There are different alternatives for representing database schemas. Besides using Data Definition Language (DDL) for relational database schemas, the XML Schema and the Web Ontology Language (OWL) are among the most popular representation mechanisms, and especially related to the increasing interest in the Semantic Web technologies. Since we focus on relational schemas in this thesis, we use the relational DDL for representing database schemas. Furthermore, for representing the integrated schemas, we use SDML, based on XML.

Since our aim is to semi-automatically match and integrate organizations' database schemas, and the relational database schemas are frequently used, the *relational schemas* constitute the main focus of our research explained as in this thesis. On the other hand, the proposed approach and the implemented SASMINT system is generalized and can be in principle extended to support other types of schemas, e.g. object-oriented as well. In the SASMINT system, relational schemas can be automatically loaded either from a relational database or from a previously saved XML file. When loading the schemas from a relational database, related metadata information, such as table and column names, is obtained from the database.

In general, the schema matching can consider different types of information as the base input, as explained in detail in Section 2.3.2. Our proposed solution however utilizes only the database schema related information (i.e. the metadata), and not the instance data. Instance data may not in general be available all the time, and using it might produce misleading and/or wrong results, if it is used alone, and without schema specification.

Different types of schema matches are addressed in Section 2.3.2. Our focus is on both simple matches (1-to-1 matches) and complex matches of type 1-to-n, n-to-1, and m-to-n.

1.5 Research Method

In the research for this thesis, we followed a method, composed of both theoretical and empirical work, as categorized in (Sørensen, 2005), which is in line with the standard scientific method.

An overview of the research phases is shown in Figure 1.3. A description of the steps in this approach is summarized below:

1. *Concept exploration and requirements analysis:* This phase constitutes the very first step of our followed method. Comprised within this step is an initial phase where an awareness of the concept of Collaborative Networks is achieved, together with an exploration of collaborative networks' data and information sharing related problems. What is further performed is an analysis of the ICT related information sharing requirements as well as the required supporting tools that would enable seamless data sharing within collaborative networks. The output of this phase is a basis, encompassing an awareness of: the collaborative networks, the data sharing requirements contained therein, the analysis of those requirements, and the resulting gap analysis, that are used for formulating the specific Research Questions that form the skeleton of our thesis work. In this step we come up with the finding that semi-automatic matching and integration of the database schemas used by the collaborating organizations is a very crucial step in solving their data interoperability/sharing related problems. We assess in this step that the resolution of this problem, i.e. the semi-automatic schema matching and integration, is definitely one main precondition to enable users with performing federated query processing transparently of its source databases over a network of collaborating entities.

2. *Identification and formulation of the Research Questions:* Based on the results of the previous step of the method, this step is where the main focus area of the thesis is devised and a context for the main research problem is established, around which the entirety of this thesis evolves. Upon an analysis of the requirements, the conceptual target area of 'schema matching and integration' is focused, and a list of research questions is built. Within the scope of the carried out research, we try to produce answers in this research for each research question listed under Section 1.2.

3. *Literature Survey and Review:* A thorough analysis of the existing theoretical and practical approaches that contribute to the resolution of the problems put forward within the context of the 'Research Questions' is performed at this step.

4. *Elaborating the Proposed Approach:* This step encompasses activities where we design a solution approach and a prototype that can be used as a supporting tool for matching and integration of heterogeneous schemas. Both the proposed approach and the prototype capitalize on the elicited data sharing requirements of collaborative networks. The scope and the objectives of our approach are refined in this step.

5. *Evaluation and also further validation of the proposed approach:* This step includes realization of the designed prototype in previous step that is used to evaluate and validate the proposed solution approach. The prototype is used to validate the adequacy of the research conducted for enabling the semi-automatic matching and integrating of database schemas from collaborative networked organizations.

Experimental evaluation of the accuracy of the proposed approach is also carried out in this step. Experiments are done using the prototype.

6. *Assessment of the results:* This phase unveils what sort of answers we have been able to produce for the research questions, and it subsumes an analysis of the answers produced. Also contained is an assessment of the value and contribution of the overall research work presented in this thesis in comparison to other related research in the area.

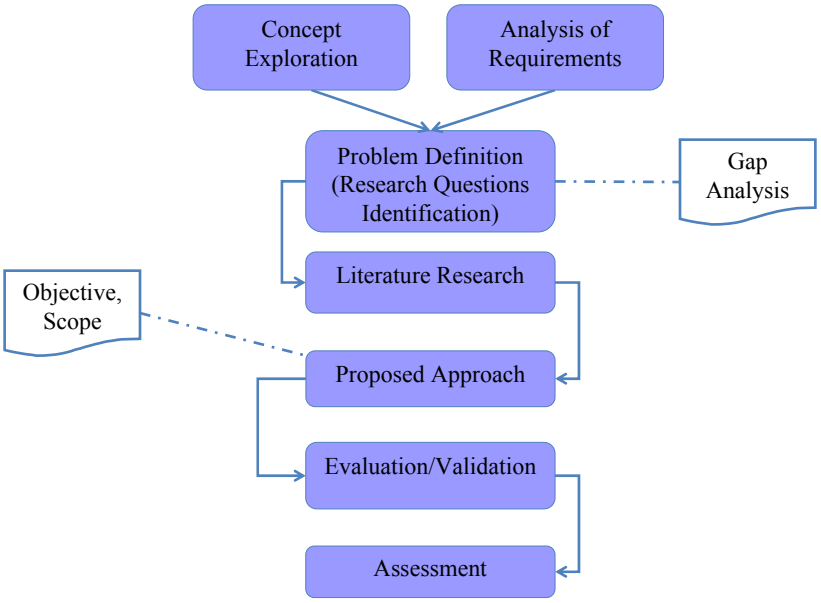

Fig. 1.3. Research method

1.6 Outline of the Dissertation

The rest of this thesis is organized as follows:

Chapter 2 provides different definitions presented in state of the art literature to refer to approaches, architectures, and systems for interlinking and/or integrating heterogeneous data provided by distributed databases in networks. Taxonomy of the terms related to an integrated information management system is provided in this chapter. Furthermore, the main features of schema matching and schema integration are addressed.

Chapter 3 aims at providing information about the heterogeneity as the most important problem to be tackled in infrastructures that enables data sharing. It addresses a number of heterogeneity (also called conflict) classifications, proposed in the literature. Furthermore, the

heterogeneity related challenges faced by the schema matching process are discussed by means of some examples.

Chapter 4 is dedicated to the SASMINT approach, proposed in the research work of this thesis. This chapter first starts with the related research, reviewing approaches focused on general database integration and interoperability, schema matching, schema integration, and ontology matching and merging. A number of open issues are addressed then to give a motivation for the proposed SASMINT approach. The rest of the chapter presents details about the phases of the SASMINT approach and how it achieves its goals.

Chapter 5 introduces the SASMINT system that is implemented to verify the approach proposed in this thesis. Details about the main components of the system are provided.

Chapter 6 provides information about the results of experimental assessment of the SASMINT system. Evaluation work covers schema matching, schema integration, as well as the Sampler components of SASMINT. Results of experiments comparing the schema matching approach of SASMINT and that of its closest competitor COMA++ are presented in this chapter.

Chapter 7 concludes the thesis with a summary of its contributions. It also presents the possible future improvements and next steps of this research.

The scientific publications related to the dissertation are listed in Appendix A.