

# Summary of the First Year Continuation Report

Lu Mao

August 28, 2008

## 1 Introduction

At the present time, data management systems need to support the explosive growth in information scale. Data managed within such systems are shared and stored across a number of heterogeneous sources, some of which are created, modelled and maintained independently by different personnel and in different ways. We call this a **dataspace**, in other words, a space of heterogeneous data sources.

Since there does not exist an universal standard to restrict the definition of schema that describes the structure of data from a particular domain (or with the same meaning), schematic heterogeneity could appear in various aspects. In other words, two schemas can represent the information from the same domain concept either use **identical** or **conflicting** schematic representations, this is called **schematic correspondences**. It can exist at the entity level; for example, one schema at a certain data source can assign an entity type named **undergraduate-student** to represent the concept: *all the undergraduate students of a particular institution*. By contrast, a schema at a different data source can assign either an *identically* named entity type, or a *differently* named entity type, for example with a shorter name **undergraduate**, to represent the same concept. Schematic correspondences can also exist at the attribute level; for example, an undergraduate student can be described by his **first-name** and **last-name** in one data source and by **full-name** and **used-name** in another data source.

One of the reasons that data management care about heterogeneity among data sources is because data can be shared and not restricted at a specific source. When user posing a query on a particular data source, his interests are not just in what information he can be retrieved from that data source, but also from other sources which can provide semantically equivalent answers for that query. Think about what web search engines like Google can serve us: they allow a user to pose a keyword on a pool of web information, and returns a set of ranked web pages with the most relevant ones on the top of the rank. However, not every time search engine can precisely identify what are the most relevant answers to match the user's desire, since it can not fully understand, and be provided with information on, *semantic relationships among heterogeneous information*. The eventual goal of data sharing and data integration is that we can get everything we want from everything there is, and the 'everything' we actually get precisely match 'everything' we should ideally get. To meet this goal, we need to know the semantic relationships between heterogeneous data sources, as well as in what ways that they correspond to each other.

Understanding the relationships between the data sources requires specifying **schema mappings**, such as one stating that **undergraduate-student** entity type in one data source corresponds to the **undergraduate** entity type in another data source under different names, and the former one has attributes **first-name** and **last-name** that correspond to the attribute, **full-name**, in the later one. Schema mappings mediate query translation among heterogeneous data sources for retrieving the most relevant answers that we can possibly get from such space of data sources. Because we are working on a large-scale of information space and the task of defining schema mappings is tedious and costly. At most time, **query evaluation** have to work on incomplete and consistent mappings to provide the most relevant results.

This research studies how to evaluate users query among heterogeneous data sources given, sometimes incomplete and inconsistent, schematic correspondences.

## 2 Research Motivation

Over the last decade, a large amount of database research effort was spent on investigating schema matching algorithms and tools. However, the output of those algorithms can only produce links between schematic elements that represent the same domain concepts (e.g.:  $Entity\_A \equiv Entity\_B$ ), with limited information describe schematic heterogeneity between data sources (e.g.:  $Entity\_A \equiv Entity\_B$  but with different name and attribute  $Entity\_A.a1$  is missing in  $Entity\_B$ ).

Anecdotal feedback received from users of commercial schema mapping tools indicates two major shortcomings when the schemas and mappings get large [Mic06]. First, visualisation and navigation tasks, such as finding *which schema elements are linked to each other*, are seriously impaired. Secondly, the task of understanding and discovering the correct semantic relationships between schema elements, provided with sophisticated schema matching techniques to generate candidate correspondences between schema elements from which the user or engineer can choose, become much harder [RDM04]. The number of possible and reasonable mappings between two data sources can be enormous [YMHF01]. To overcome the first problems, advanced visualisation techniques [RCC05] have been proposed to improve usability in such situation. To overcome the second issue, techniques have been proposed for debugging and refining schema mappings, such as [CT06, YMHF01] and for inferring structural relationships from instances [GW97].

Referring to the vision of dataspace [FHM05, HFM06], schema mappings should be automatically inferred and are almost never sufficient and consistent, therefore the management of dataspace has to support query evaluation in the context of partial and inconsistent mappings. Nevertheless, existing schema mapping techniques, tools and algorithms should be acknowledged for mapping construction and refinement in dataspace. However, those solutions are still incomplete.

This research work continues on the contribution made by kim et al on the identification, classification and resolution of schematic heterogeneity [KS91, KCGS93]. As mentioned early, this early research work can only serve as a guide for manual construction (by database administrators) of a multidatabase

[BDK<sup>+</sup>88, BR90] which is an integration of heterogeneous source databases. The manual construction process consist of, firstly, identifying the classified schematic conflicts among the sources. Secondly, each identified conflict is resolved based on a set of corresponding techqniues classified in [KCGS93]. Finally, a mapping is defined in the form of a query which specifies how to populate (materialise) a multidatabase by data stored in source databases.

In this research, we would like to extend kim's work by one step further to describe and model schematic correspondences between heterogeneous data sources in the context of dataspace. Assume that information on a set of schematic correspondences between heterogeneous data sources in a dataspace are given, we study how to leverage these information and kim's work on schematic heterogeneity resolution techniques to develop automated mechanisms for generating query views (mappings), and for evaluating queries among heterogeneous data sources.

The technical contributions to be made at the end of this exploration are intended to serve as a key support for pay-as-you-go data management systems.

- *Contribution 1*: To refined the classification of schematic correspondences (in Chapter 3) originally conducted by Kim et al in [KS91, KCGS93] by:
  1. identifying more specific types of schematic correspondences.
  2. modelling and documenting correspondences among schematic elements in the context of dataspace.

This contribution can influence the work on creating low cost query view generation mechanisms and information integration.

- *Contribution 2*: design automated mechanisms in the form of algorithms to automatically generate query views (Chapter 4) and evaluate user queries (future work) posed on heterogeneous data sources influenced by *Contribution 1* and manual schematic resolution techniques proposed in [KS91, KCGS93].

### 3 Research Objectives

In order to meet the research goal, the following objectives need to be achieved:

1. To study and refine the classification of schematic heterogeneity proposed by Kim [KS91, KCGS93].
2. To study existing techqniues on query reformulation [CLL01] in the field of data integration [Len02, HRO06].
3. To produce documentation to describe schematic correspondences between heterogeneoue sources in a dataspace.
4. To develop algorithms for generating query views between pairs of heterogeneous sources informed by pre-given schematic correspondences. Each view specifies how schematic elements, that represent some data in one source, correspond to elements in the other source.

5. To develop automated query evaluation mechanisms mediated by query views that translate queries between schematic corresponding sources.
6. To study how to detect and minimise (in other word, to clean) the incompleteness and inconsistency on schematic correspondences in the context of a large-scale and heterogeneous data sources.
7. To study how to minimise the impact on query evaluation in the context of partial and inconsistent schematic correspondences.
8. To conduct experimental studies to evaluate the work on view generation and query evaluation.

## 4 Research Progress

The results obtained up to the current progress are the following:

1. A good knowledge on the area of data integration, dataspace, query reformulation on heterogeneous data and schematic correspondence has been acquired and a thorough literature review on these fields has been carried out.
2. The initial research achievement is on the documentation of schematic correspondences early classified by kim in [KCGS93]. The documentation describes and explains regarding characteristics of each schematic correspondence, such as its participating schematic elements, its type, the certainty of this correspondence, how to resolve a conflict and other correspondence specific informations.
3. Developed algorithms, driven by information on schematic correspondences described in (3), to generate query views (schema mappings) that describe the semantic relationships between pairs of schematic elements. This work is currently restricted to some types schematic correspondences, which will be refined in future works.

## 5 Future Work

The work described in this report laid the foundation for future research on query evaluation on heterogeneous sources that makes use of schematic correspondences. The next step will achieve the following aims:

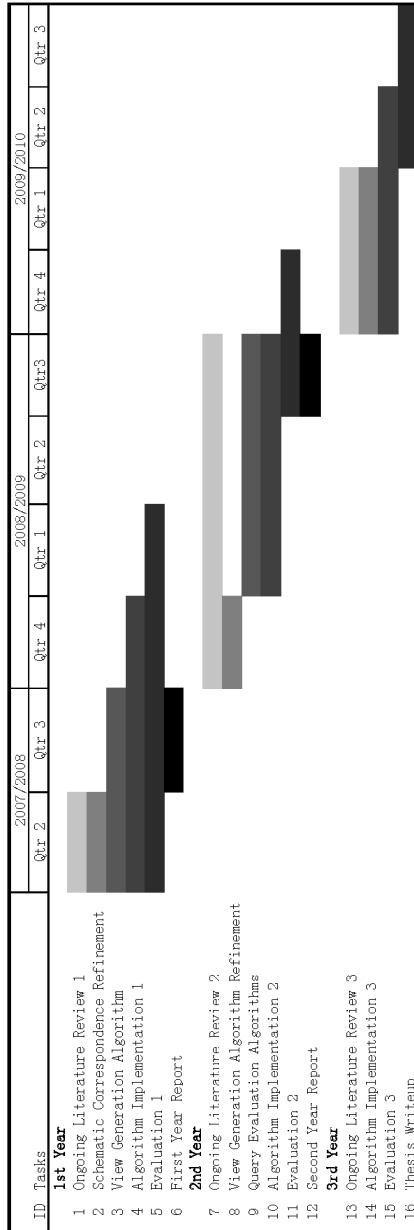
- Refining the view generation algorithms from its current state by to be capable of handling more types of schematic correspondences, such as many-to-many entity correspondence with horizontal partitioning and many-to-many attribute correspondences.
- Design mechanisms for evaluating queries on generated mappings by studying existing approaches from query reformulation in data integration (e.g. GAV and LAV).

- Since the scale of data in a dataspace is increasingly large and users are not skilled enough to provide intact and consistent schematic correspondences, it is impossible to specify precise mappings between heterogeneous data. Both the mapping generation and query evaluation on schematic correspondences has to incorporate consideration on and means to coop with this fact.
  - To identify in what ways that schematic correspondences can be incomplete and inconsistent, e.g. two many-to-many entity or attribute correspondences can produce inconsistency when the source and target element set of one correspondence is included in another correspondence.
  - To classify inconsistent correspondences, e.g. into conflicting correspondences, redundant correspondences, overlapping correspondences, and etc.
  - Design mechanisms to detect incompleteness and inconsistency in correspondences.
  - Design mechanisms to eliminate or minimise impact of incompleteness and inconsistency on query evaluation
- Evaluating mapping generation and query evaluation on large scale and heterogeneous real world datasets.

The future research plan is outlined in Figure 1 as a Gantt chart .

## 6 Anticipated Thesis Structure

1. Introduction
  - Data Integration
  - Dataspace
  - Schematic Correspondence
  - Query Reformulation
  - Motivation and Objectives
  - Thesis Structure
2. Information Management
  - Data Management Architecture
  - Dataspaces
3. Related Work
  - Data Integration Architectures
  - Query Reformulation in Data Integration
  - Pay-as-you-go Data Management
  - Schematic Correspondence Classification
  - Summary



6  
Figure 1: A Gantt chart outlining the timing for future research plan.

4. Schematic Correspondence in Heterogeneous Data Sources
  - Schematic Correspondence Modelling
  - Schematic Correspondence Documentation
  - Summary
5. Query Evaluation on Schematic Correspondences
  - Query View Generation
  - Inconsistency in Schematic Correspondences
  - Query Evaluation Algorithms
  - Summary
6. Experimental Studies
  - Experiment Setup
  - Methodology
  - Result Analysis
  - Summary
7. Conclusion and Future Work
  - Research Achievement
  - Significance of Contributions
  - Future Works

## References

- [BDK<sup>+</sup>88] V. Belcastro, A. Dutkowski, W. Kaminski, M. Kowalewski, C. L. Mallamaci, S. Meszyk, Tommaso Mostardi, F. P. Scrocco, Witold Staniszki, and G. Turco. An overview of the distributed query system dqs. In *EDBT '88: Proceedings of the International Conference on Extending Database Technology*, pages 170–189, London, UK, 1988. Springer-Verlag.
- [BR90] P. Bodorik and J. S. Riordon. System integration in multidatabases. In *SIGSMALL '90: Proceedings of the 1990 ACM SIGSMALL/PC symposium on Small systems*, pages 160–163, New York, NY, USA, 1990. ACM.
- [CLL01] D. Calvanese, D. Lembo, and M. Lenerini. Survey on methods for query rewriting and query answering using views, 2001.
- [CT06] Laura Chiticariu and Wang-Chiew Tan. Debugging schema mappings with routes. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 79–90. VLDB Endowment, 2006.
- [FHM05] Michael Franklin, Alon Halevy, and David Maier. From databases to dataspace: a new abstraction for information management. *SIGMOD Rec.*, 34(4):27–33, 2005.

- [GW97] Roy Goldman and Jennifer Widom. Dataguides: Enabling query formulation and optimization in semistructured databases. In Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederick H. Lochovsky, Pericles Loucopoulos, and Manfred A. Jeusfeld, editors, *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases*, pages 436–445. Morgan Kaufmann, 1997.
- [HFM06] Alon Halevy, Michael Franklin, and David Maier. Principles of dataspace systems. In *PODS '06: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–9, New York, NY, USA, 2006. ACM.
- [HRO06] Alon Halevy, Anand Rajaraman, and Joann Ordille. Data integration: the teenage years. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 9–16. VLDB Endowment, 2006.
- [KCGS93] Won Kim, Injun Choi, Sunit Gala, and Mark Scheevel. On resolving schematic heterogeneity in multidatabase systems. *Distrib. Parallel Databases*, 1(3):251–279, 1993.
- [KS91] Won Kim and Jungyun Seo. Classifying schematic and data heterogeneity in multidatabase systems. *Computer*, 24(12):12–18, 1991.
- [Len02] Maurizio Lenzerini. Data integration: a theoretical perspective. In *PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, New York, NY, USA, 2002. ACM.
- [Mic06] Philip Bernstein Microsoft. Incremental schema matching, 2006.
- [RCC05] George G. Robertson, Mary P. Czerwinski, and John E. Churchill. Visualization of mappings between schemas. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 431–439, New York, NY, USA, 2005. ACM.
- [RDM04] Erhard Rahm, Hong-Hai Do, and Sabine Massmann. Matching large xml schemas. *SIGMOD Rec.*, 33(4):26–31, 2004.
- [YMHF01] Ling Ling Yan, Renée J. Miller, Laura M. Haas, and Ronald Fagin. Data-driven understanding and refinement of schema mappings. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 30(2):485–496, 2001.