Schema Matching and Mapping: From Usage to Evaluation

Angela Bonifati
CNR, Italy
bonifati@icar.cnr.it

Yannis Velegrakis
University of Trento, Italy
velgias@disi.unitn.eu

ABSTRACT

This tutorial provides an overview of current evaluation techniques for schema matching and mapping tasks and tools, alongside existing and broadly used evaluation scenarios. The objective is to introduce the audience into the area of matching and mapping system evaluation, and to highlight the need for leveraging robust benchmarks and yardsticks for the comparison of the different matching and mapping tasks. Open research problems will be identified and presented. The tutorial is for both experienced researchers and unfamiliar investigators looking for a quick and complete introduction to the topic.

Categories and Subject Descriptors

H.2.5 [**Database Management**]: Heterogeneous Databases—*Data Translation*

General Terms

Measurement

Keywords

Schema matching, schema mapping, data exchange, data integration, evaluation

1. INTRODUCTION

Data heterogeneity has always been prevalent across repositories and information systems. The data management research community has been working toward tackling it for several decades. To cope with heterogeneity and ensure the interoperability of the underlying systems, a fundamental requirement is the identification of structures in two different schemas representing the same real world entity or business artifact, and generating expressions that specify exactly how these structures relate to each other and can translate data conforming to the first structure into data conforming to the second. These two tasks are found in the literature under the names of matching [1] and mapping [2], respectively.

Although matching and mapping are often used as interchangeable terms, they actually identify two complementary activities that are indeed becoming pervasive in our daily life: *(i)* in data integration [3], matching and mapping are fundamental components used to express the relationship between the local and the mediated schema; *(ii)* in data exchange scenarios [4], they embody the transformation from the source format into a target representation; *(iii)* in schema evolution and maintenance [5], they represent the connection between old and new version of the schemas and show how the old instances are to be converted into the new representation; and *(iv)* in web applications, modern service-oriented facilities and data publishing tools, they regulate the translation from the native data model into the data model used for communication with other applications and/or used for the publishing of data. Moreover, mappings have recently been studied even for stream schema applications [6], or in P2P distribution paradigms [7].

Matching and mapping had been tasks traditionally performed manually by a data designer having a good understanding of the semantics of the two schemas and a good knowledge of the language used to express the transformations. Unfortunately, the tremendous increase in the complexity and size of modern schemas and the intricate semantics of their relationships made the task laborious, time consuming, and error-prone. Tool support turned to be a fundamental need. Currently, there is a plethora of commercial mapping systems available in the market [8, 9, 10] and research prototypes [7, 11, 12, 13].

2. RATIONALE, MOTIVATION AND CHALLENGES

Despite the large number of mapping systems that are currently available, there has been no generally accepted benchmark or technique for evaluating and comparing these tools. The existence of such a benchmark is of major importance for assessing the relative merits of the systems. It can help customers in making the right investment decisions by selecting among the different tools those that better fit their business needs. It can offer developers a platform to stress test their systems, compare with competitors and highlight limitations, with all these boosting competition and serving as a driving force toward mapping systems of better quality and with more services offered. A recent workshop on Information Integration had the lack of benchmarks as one of its main topics for discussion [14].

Unfortunately, although new matching and mapping tools are daily becoming available to a broader audience, there has been some form of confusion regarding their exact nature, goals, core functionalities, expected features and capabilities. Above all, this has made the discovery, design and use of performance measurements for such systems a challenging task, undermining efforts for the

EDBT 2011, March 22–24, 2011, Uppsala, Sweden.

creation of a globally accepted benchmark. For instance, it is not clear for a mapping system what is the input language that the mapping designer uses to describe the specifications for the mappings she is willing to design. Such a language can be either as simple as direct 1-1 lines between schema elements, or very complex such as an XSLT script or an XQuery. Any assumption for the input language makes unfair the comparison to other systems that do not operate under such assumption.

Although the scientific literature contains a large number of evaluation metrics that have been proposed at different times, there is no systematic work that consolidates them under one unified framework. For matching systems there is already a respectable amount of work done [15], but for mapping systems, existing efforts toward a systematic design of a benchmark [16, 11] are in their infancy.

The need for design and development of globally accepted comparison standards is becoming more apparent than ever before. Awareness needs to be built among researchers and motivations be given for working toward that direction. On the other hand, researchers and practitioners have to learn to use generally accepted evaluation methods when performing an experiment evaluation of their system against competitors. Before initiating any efforts toward this direction, it is fundamental for the research community to obtain a good understanding of the nature, features, goals and capabilities of matching and mapping systems.

3. OBJECTIVES

The objective of this tutorial is three-fold. First, to educate the audience on the roles and functionalities of the matching and mapping systems. It intends to provide a generic overview of the basic principles on which these systems are designed and on the assumptions upon which they are build. The second objective is to provide a list of the expected features for a benchmark for such systems and highlight the reasons that make its design more challenging than other benchmarks such as those of traditional query engines. The final objective is to provide a complete, unified and systematic presentation of all the efforts that have been proposed so far on evaluating the matching and mapping process.

4. SCOPE AND DEPTH

The tutorial covers all the concepts of the spectrum of design, development and use of matching and mapping systems. Formal definitions are provided alongside intuitive explanations, to ensure that the audience has the required background. While illustrating schema matching and mapping tasks, the challenging issues involved in the correct evaluation of the existing tools will be highlighted. Metrics that have been proposed in the literature for evaluating the different aspects of matching and mapping will be formally defined, intuitively explained and demonstrated through their application to common matching and mapping tasks. The final goal is to let the different kind of users evaluate the usefulness of the different evaluation criteria and decide which to adopt when evaluating their tools for a specific task at hand. Finally, a detailed list of existing efforts on benchmarking schema matching and mapping tasks will be provided, and for each one, a comprehensive description of its advantages and disadvantages will be discussed. During the tutorial, open issues and research problems will be identified and analyzed.

5. BRIEF OUTLINE

The presenters will start by giving an overview of the schema matching and mapping problems, and show the main components of data translation and integration architectures. They will also highlight the difficulties of the above problems, and the expectations that developers and practitioners often have with respect to the solutions. Then, the challenges of evaluating the matching and mapping tools will be underlined, alongside real world and synthetic scenarios that these tools may advocate. Finally, the presenters will discuss the various quality metrics that are needed for a sound evaluation, including (but not limited to) the efficiency, effectiveness and the human intervention. A more detailed outline of the tutorial is the following:

- Introduction
- The Matching and Mapping problems
- Design, Development and Use of Matching and Mapping Tools and Techniques
- Challenges in Matching and Mapping System Evaluation
- Collecting Good Real World Scenarios for Testing
- Systematic Generation of Synthetic Scenarios
- Metrics Measuring Efficiency
  - Matching and Mapping Generation Time
  - Data Translation Performance
  - Human Effort
- Metrics Measuring Effectiveness
  - Number of Supported Scenarios
  - Quality of the generated Matchings/Mappings
  - Quality of the generated Target Instance
  - Quality of the generated Target Schema
  - Conformance to data examples
- Conclusion

6. TARGET AUDIENCE

The tutorial aims at a broad range of researchers, students, IT professionals and practitioners, and developers. Anyone working in information integration, data exchange, data management, benchmarking, experimental evaluations, or other related fields, will benefit from this tutorial. *Students* and researchers will not only get a good introduction to the topic with a complete coverage of the stat-of-the-art, but will also find a number of challenging research problems in these emerging technologies on which they may decide to focus their future research efforts. *Practitioners* will get a good overview of the benefits that matching and mapping systems can offer nowadays and learn how they can use this to improve the productivity of their businesses. They will also learn how to evaluate the mapping products that are currently available in the market, and how to choose those that can more successfully execute a particular task at hand, or better fit their general needs. *Developers* of mapping and matching tools will learn how their tools can be evaluated and how they can compare them with competitors' products. Such an evaluation will allow them to identify limitations and will promote the development of new better mapping and matching tools. The tutorial itself will also offer the developers a number of ideas on how to improve their existing products.

7. PREREQUISITES

The structure of the tutorial has been carefully designed in order to accommodate time for providing all the required background knowledge. This has been deliberately done in order to bring into context attendees unfamiliar with the topic, and to provide a shared terminology and common understanding of the basic concepts among the experienced researchers, since the latter may see different aspects of the matching and mapping problem and may have different views and definitions for the various concepts and goals.

8. SHORT BIOGRAPHIES

Angela Bonifati is a researcher in Computer Science at CNR, Italy since 2002 and has been qualified as Associate Professor at the University of Basilicata in 2010. She received her Ph.D. from Politecnico di Milano in 2002. Her research interests are on the interplay of structured and unstructured data, query and update languages for XML, schema matching and mapping, query optimization and distributed databases. She has participated in two projects on schema mapping and query translation [17, 7]. She has co-edited a book entitled 'Schema Matching and Mapping' with Z. Bellahsene and E. Rahm. The book is part of the Springer-Verlag Series on 'Data-centric Systems and Applications' and is scheduled to appear at the beginning of 2011. She has spent time as an invited researcher at Inria (France) and the University of British Columbia (Canada). She holds a US patent, has served in many program committees of international database conferences and has been Vice Chair of ICDE 2011, for the Semistructured Data, XML and Web Data Management track.

Yannis Velegrakis is a faculty member at the University of Trento, with expertise in schema mapping, interoperability, heterogeneous information integration, data exchange, view management, and keyword searching. He graduated from the University of Toronto, with a thesis on mapping management. Prior to joining the University of Trento, he held a researcher position at ATT Research Labs (USA). He has also spent time as a visitor at the IBM Almaden Research Center (USA), where he participated in the development of the Clio schema mapping took, the Center of Advanced Studies of the IBM Toronto Lab (Canada), and the University of California, Santa-Cruz (USA), where he and his collaborators developed the STBenchmark, a generic and multi-purpose benchmark for schema mapping systems. He has served in program committees of many national and international conferences, has been a reviewer for numerous international journals and was a Marie Curie Reintegration fellow between 2006 and 2008. He has been a general chair for the DESWeb 2010 and 2011 ICDE Workshops and will be a general chair of VLDB 2013.

9. REFERENCES

[1] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching," *VLDB Journal*, vol. 10, no. 4, pp. 334–350, 2001.

[2] R. J. Miller, L. M. Haas, and M. A. Hernandez, "Schema Mapping as Query Discovery," in *VLDB*, 2000, pp. 77–88.

[3] M. Lenzerini, "Data Integration: A Theoretical Perspective," in *PODS*, 2002, pp. 233–246.

[4] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa, "Data exchange: semantics and query answering," *Theoretical Computer Science*, vol. 336, no. 1, pp. 89–124, 2005.

[5] B. S. Lerner, "A Model for Compound Type Changes Encountered in Schema Evolution," *ACM TODS*, vol. 25, no. 1, pp. 83–127, Mar. 2000.

[6] P. M. Fischer, K. S. Esmaili, and R. J. Miller, "Stream Schema: Providing and Exploiting Static Metadata for Data Stream Processing," pp. 207–218, 2010.

[7] A. Bonifati, E. Q. Chang, T. Ho, L. V. S. Lakshmanan, R. Pottinger, and Y. Chung, "Schema mapping and query translation in heterogeneous P2P XML databases," *VLDB Journal*, vol. 19, no. 2, pp. 231–256, 2010.

[8] Stylus Studio, "XML Enterprise Suite," 2010, www.stylusstudio.com.

[9] Altova, "MapForce," 2008, http://www.altova.com.

[10] Microsoft, "Visual Studio," 2005, msdn2.microsoft.com/en-us/ie/bb188238.aspx.

[11] A. Bonifati, G. Mecca, A. Pappalardo, S. Raunich, and G.Summa, "Schema Mapping Verification: The Spicy Way," in *EDBT*, 2008, pp. 85 – 96.

[12] L. Popa, Y. Velegrakis, R. J. Miller, M. A. Hernandez, and R. Fagin, "Translating Web Data," in *VLDB*, Aug. 2002, pp. 598–609.

[13] G. H. L. Fletcher and C. M. Wyss, "Data Mapping as Search," in *EDBT*, 2006, pp. 95–111.

[14] editor, Ed., *Bertinoro Workshop on Information Integration*, org. publisher, 2007, www.dis.uniroma1.it/~lenzerin/INFINT2007.

[15] J. Euzenat and P. Shvaiko, *Ontology matching*. Heidelberg (DE): Springer-Verlag, 2007.

[16] B. Alexe, W. C. Tan, and Y. Velegrakis, "STBenchmark: towards a benchmark for mapping systems," *Proc. of VLDB Journal*, vol. 1, no. 1, pp. 230–244, 2008.

[17] A. Bonifati, G. Mecca, A. Pappalardo, S. Raunich, and G. Summa, "The Spicy system: towards a notion of mapping quality," in *SIGMOD*, 2008, pp. 1289–1294.

[18] R. Fagin, P. G. Kolaitis, and L. Popa, "Data exchange: getting to the core," in *PODS*, 2003, pp. 90–101.

[19] B. Alexe, W. C. Tan, and Y. Velegrakis, "Comparing and evaluating mapping systems with STBenchmark," *Proc. of VLDB Journal*, vol. 1, no. 2, pp. 1468–1471, 2008.