

# $\mu$ BE: User Guided Source Selection and Schema Mediation for Internet Scale Data Integration

Ashraf Aboulnaga

Kareem El Gebaly

Daniel Wong

*University of Waterloo*  
{ashraf, kelgebal, d5wong}@cs.uwaterloo.ca

## 1. Introduction

The typical approach to data integration is to start by defining a common mediated schema for all the data sources, then to match and map the data sources to this mediated schema. This assumes that the user (1) knows all the sources that should be included in the data integration system, and (2) knows all the concepts that are expressed in these data sources (which will come together to form the mediated schema). These assumptions do not necessarily hold in cases where the data integration task involves hundreds or thousands of data sources, especially if these data sources are new to the user. Such “Internet-scale” data integration tasks are becoming increasingly more common, arising in scenarios such as Web data integration, peer-to-peer networks supporting structured queries, personal information management, and scientific data management.

In these Internet-scale data integration tasks, the user most likely has an understanding of the important concepts in the domain they are working in, but may not know all the different concepts expressed in all available data sources, and how they are expressed in these sources. Thus, it is difficult for the user to decide a priori on the mediated schema. Moreover, the user may not want to include all available data sources in the data integration system being defined.

We can see that a user who is building an Internet-scale data integration system needs to solve two problems: (1) which sources to include in the data integration system, and (2) what mediated schema to use. These decisions should be based on the schemas of the data sources, the quality and quantity of the data at the different sources, and other source characteristics such as reputation or reliability. An effective way to make these decisions is to *iteratively explore* the space of possible choices to find a good solution. While exploring the space, the user gains more knowledge of the problem domain and can better define the desired solution. We demonstrate a data integration tool that helps users in this iterative exploration task. We call our tool  $\mu$ BE, for *Matching By Example*. Details about  $\mu$ BE are given in [1].

The goal of  $\mu$ BE is to choose a set of data sources and a global mediated schema over these sources. A mediated

schema in  $\mu$ BE consists of a set of *Global Attributes* (GAs). A GA is a set of attributes from different sources expressing the same concept. Finding the GAs representing the best mediated schema is at the core of  $\mu$ BE.  $\mu$ BE formulates this task as a constrained optimization problem, where the objective function is to maximize the overall quality of the chosen mediated schema, subject to user-specified constraints. Users can specify *source constraints*, which require certain sources to be part of the solution, without specifying how the schemas of these sources should be matched with other sources. Users can also specify *GA constraints*, which require that certain GAs be part of the chosen mediated schema. The source and GA constraints specified by the user guide  $\mu$ BE in its search of the space of possible solutions. After  $\mu$ BE finds a solution, the user can modify the specification of the optimization problem and request a new iteration of  $\mu$ BE. This iterative process continues until a solution is found that is acceptable to the user.

$\mu$ BE is implemented as two separate modules: a *Solution Engine* and a *User Interface*. The Solution Engine is concerned with solving the optimization problem described above. The User Interface allows users to modify the optimization problem between iterations, and to analyze the details of the solution chosen by  $\mu$ BE. The  $\mu$ BE User Interface and the interactions it allows with users are the focus of this demo. The interface consists of two components: the *Solution Finder*, and the *Solution Analyzer*.

## 2. Solution Finder

The Solution Finder (SF) allows users to iteratively solve a sequence of optimization problems to find the desired data integration system (Figure 1). The interaction with the Solution Engine is through XML files, with the same XML schema used for both input and output. The SF allows users to manipulate these XML files to modify constraints, quality evaluation functions (QEF), or weights on the QEFs.

To modify the GA and source constraints, the SF takes full advantage of our design decision of using the same GA definition for input constraints and attributes of the output mediated schema. The GA Constraint Editor allows users to add GA constraints by choosing attributes from all available sources, or from the set of sources currently selected by the

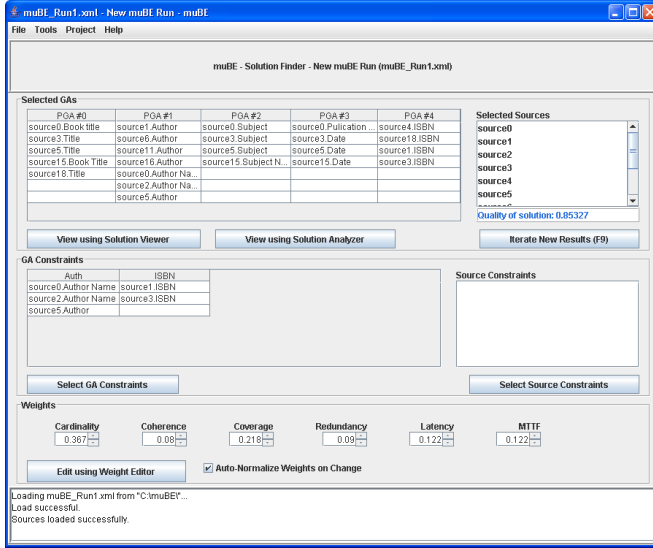


Figure 1. Solution Finder.

Solution Engine (Figure 2). A weight editor allows the user to modify the weights of QEFs and add new QEFs.

### 3. Solution Analyzer

To define the constraints, weights, and QEFs for the next iteration of  $\mu$ BE, the user needs a better understanding of the current solution. The Solution Analyzer (SA) provides two views on the current solution that allow the user to drill down and examine why this solution was chosen.

The first view is the Candidate Viewer (Figure 3). Here, the user can see the solution with the highest overall quality along with the few solutions with the next highest overall quality. These are all strong candidates that contain useful information. The overall quality of each candidate solution is broken down into different QEFs, which allows the user to see the effect of the current weights. The user can examine the data sources and GAs of any candidate solution, and can use them to define new source and GA constraints. The user can also modify the weights of the QEFs and examine the effect that this has on the solution.

The second view defined by the SA is the Source Contribution Viewer (Figure 4), which allows the user to drill down even more on a candidate solution and examine the contribution of each data source to each QEF in this solution. If a data source is found to be desirable, the user can add it as a source constraint for the next iteration of  $\mu$ BE.

Without the SA, setting weights and constraints would be a difficult and opaque task. The SA helps significantly in this task by allowing the user to understand the detailed characteristics of the solution chosen by  $\mu$ BE.

### References

[1] A. Aboulnaga and K. El Gebaly.  $\mu$ BE: User guided source selection and schema mediation for internet scale data integration. In *Proc. IEEE Int. Conf. on Data Engineering*, 2007.

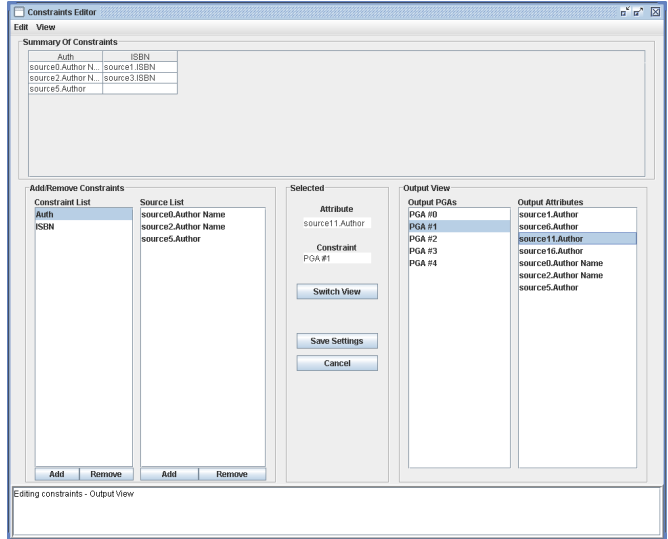


Figure 2. GA Constraint Editor.

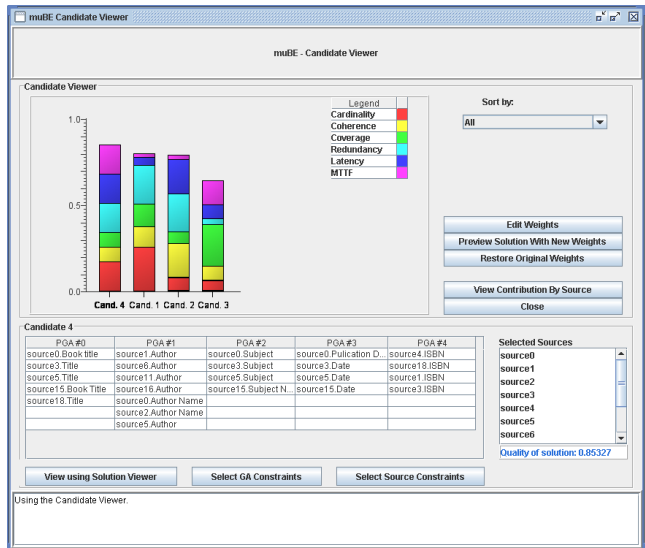


Figure 3. Candidate Viewer.

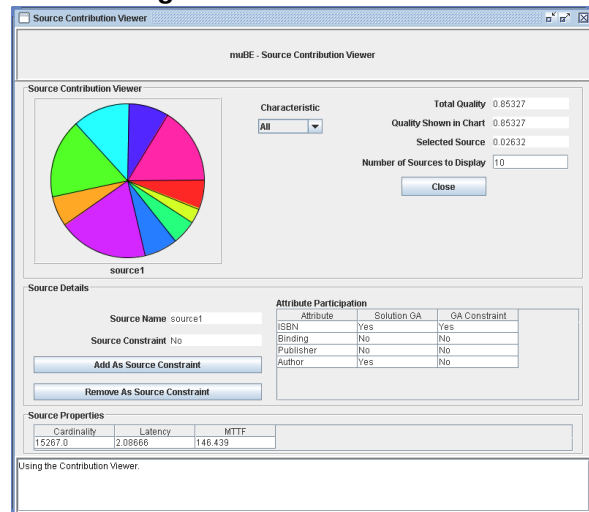


Figure 4. Source Contribution Viewer.