A Model for Semantic Equivalence Discovery for Harmonizing Master Data

Baba Piprani

MetaGlobal Systems, Canada
babap@attglobal.net

Abstract. IT projects often face the challenge of harmonizing metadata and data so as to have a "single" version of the truth. Determining equivalency of multiple data instances against the given type, or set of types, is mandatory in establishing master data legitimacy in a data set that contains multiple incarnations of instances belonging to the same semantic data record . The results of a real-life application define how measuring criteria and equivalence path determination were established via a set of "probes" in conjunction with a score-card approach. There is a need for a suite of supporting models to help determine master data equivalency towards entity resolution---including mapping models, transform models, selection models, match models, an audit and control model, a scorecard model, a rating model. An ORM schema defines the set of supporting models along with their incarnation into an attribute based model as implemented in an RDBMS.

Keywords: Entity Resolution, Master Data, Semantic Equivalence, semantic interoperability, data equivalency.

1 Data Redundancy and Integration Issues

Data duplication and data integrity are major issues confronting IT applications today. It is not uncommon to come across multiple sets of redundant data on customers and other items of primary focus in an organization. A good average figure that is the norm is that there is about 30 to 60 percent redundant data or duplicated data on clients or customers in a typical organization. The search goes on ….for the 'single version' of the truth.

Enter Master Data Management, which basically refers to non-transactional data, or reference data. There are basically 2 kinds of reference data---the common reference data like types and categories of data pertaining to the properties or characteristics of data, and the other being 'master file' type data like, customers, clients, vendors, products etc…essentially data that the organization uses for tracking through transactions.

Both of these types of reference data need to be harmonized and coordinated. Many events occur during the life cycle of an organization that causes data to ride the redundancy paddy wagon. Mergers, acquisitions, or even simple lack of governance controls, poorly defined business processes etc. contribute towards multiple copies of client data, creating exponentially increasing issues of data integration.

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 649–658, 2009.

This paper addresses a model for the solution steps to be taken that are involved in integrating or harmonizing such data, as implemented in an on-going application shown as a use case.

2 Background of Use Case

A critical air traffic situation arose in Canada on September 11, 2001 where hundreds of planes bound to the US over the Atlantic were re-routed to Canada. There was an urgent flurry of activity to determine the most suitable and critical match between airport runaway properties; emergency service facilities; airside airport services; passenger airport facilities etc. and, the incoming aircraft with respect to the aircraft type; aircraft size; number of passengers in each aircraft; fuel conditions in each aircraft etc. This presented a Herculean task for Transport Canada (TC), the regulatory agency responsible for the transportation sector in Canada, to access such data across multiple applications on demand! Needless to say, not all applications were in a condition to be cross-navigated considering the urgent timing of the requirement.

There was lack of navigability across the various involved systems for items like Airport locations, aircraft types, cities within Canada, etc. In other words, each application and database has its own native referencing scheme that did not permit establishing of cross-walks at the metadata nor the value based level. Some underlying issues:

- Mappings were not always available (no total 1:1). In other words, not all airport locations had ICAO codes or IATA codes, since Canada has CFS (Canadian Flight Supplement) codes for the domestic airports.
- Inconsistent identification and different terminology for referencing the same concept.
- Inconsistent identification of the different types of aircraft.
- rounding errors for latitude/longitude produced multiple airports at exact same location
- Cities within Canada were inconsistently referenced across multiple databases….and so on.

The Transport Canada - Transport Object Dictionary (TOD) project was born to establish a cross-walk across the various applications so that a virtual access approach is feasible by establishing common mappings across various applications and using these common mappings to automatically access data across applications thus simulating a virtual on-demand data warehouse [1] so as to be able to conduct Business Intelligence analysis scenarios.

In order to establish this, one of the first steps was to harmonize data based on selected main concepts---Location, Carrier Organizations, Carrier Individuals, and Aircraft etc.

3 Establishing Semantic Mapping - An ORM Mapping Model (Type Level)

One of the first steps in the semantic equivalency exercise is to determine the source and target mappings. As explained in [1], in the absence of one finite or global

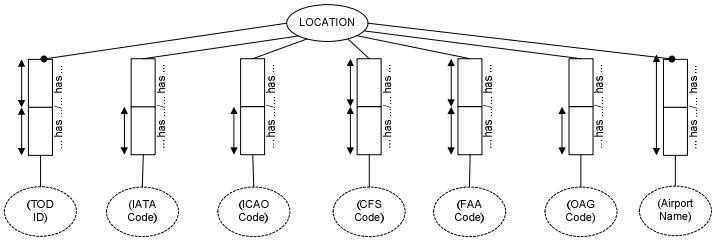


Fig. 1. Mapping to a global identifier

identifier for a defined location, a global identifier was established and the mappings were done from the source to the target global identifier only, see Fig. 1.

There are several papers written in the area of semantic interoperability and entity resolution see [10-15] and works by Embley and Ullman. Most of the literature deals with schema mapping and wrappers using semi-automated means to establish mappings. However, in practice, the lack of data quality, the inconsistency of data, the lack of standard formats, typographical misspellings, typo errors, non-standard notations say for organizations, colloquial names, abbreviations, whitespaces etc., cause serious matching problems in de-duplication or merging for entity resolution. More often than not, there is a serious lag or gap of enterprises in adopting the industry accepted practices for coding or naming is not followed, thus creating a plethora of variants within the same business community. The common message that comes across is that this integration step of mapping and merging for entity resolution is not a trivial task and there is much work ahead.

Mapping across data attributes needs to be harmonized at 2 levels---the metadata level (or type) and at the value level (instance), beginning with the type level and to progress the implementation to the instance level.

The TOD project first developed a metadata cross-walk across multiple applications determining the best fit for the metadata mappings using the accompanying ORM Mapping Model as in Fig. 2.

The ORM Mapping Model shows mappings between a source data element in the form of a qualified column being mapped to a target qualified column. The model addresses renaming, conversion, filtering, transformation and introduction of a global identifier. Only a partial model is shown that is relevant to the mappings. Other parts of the model contain the mapping and referencing to the data element as defined in the ISO Metadata Registries standard [2], which is not shown.

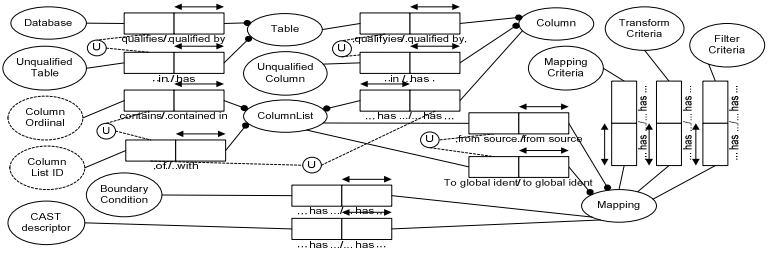


Fig. 2. ORM Mapping Model (type level)

As ide from the source-target (from-to) mappings based on a column list that represents one or more qualified columns, the mapping model also uses other mapping sub-models: Boundary Conditions ; CAST descriptor; Mapping Criteria; Transform Criteria; and Filter Criteria. These sub-models contain criteria that are effective for the global set of values at the metadata level and not against operations on an individual row set.

Boundary Conditions: denotes range, or value or other integrity constraint on the involved column(s) for mapping purposes to be included in the data mapping, usually denoted and exercised by an SQL CHECK clause or User Defined Function. *e.g. The carrier UID in CPF is CHECK (carrierUID BETWEEN 1 AND 999999).* Contains the actual CHECK clause to be used that can be automatically included in a dynamic SQL statement.

CAST descriptor: Any data type conversions on a global scale *(*see later section for value based CASTing), usually exercised via an SQL CAST predicate. Contains the actual CAST predicate to be used that can be automatically included in a dynamic SQL statement.

Mapping Criteria: Any matching criteria beyond the involved columns that affect the mapping *e.g. where type=xx, colour = yy etc*.

Transform Criteria: Any conversion or transform that includes a change of variables or coordinates in which a function of new variables or coordinates is substituted for each original variable, *e.g. old Address Type = 2, becomes new Address Type = Business.* Contains the actual User Defined Function or code snippet to be used that can be automatically included in a dynamic SQL statement.

Filter Criteria: Any applicable controlling criteria to limit the values of the selected set. Contains the actual User Defined Function or code snippet to be used that can be automatically included in a dynamic SQL statement.

4 Establishing Semantic Mapping - An ORM Mapping Model (Instance Level)

The ORM Mapping Model at the type level is useful as the meta driver for any Extract Transform Load utilities and for generating dynamic SQL statements to drive the data transfer or migration based on the established mappings.

Taking the ORM Mapping Model to the instance level, the ORM Mapping Model (instance level) shows the exact transforms to be used for a given local row level instance, as seen in Fig. 3.

The ORM Mapping Model at the instance level is used to cross-correlate or establish concordance between values from one application system to another. An example would be if the row of data pertaining to Ottawa Airport Location denoted by IATA code YOW has a primary key column <airport_reporting> in Application A with value 1111, which is mapped to Application B which has a primary key column <primary airport id> with value 2259, the ORM Mapping Model (type level) will

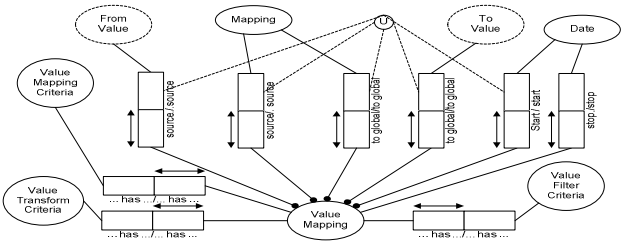


Fig. 3. ORM Mapping Model (instance level)

contain the mappings between Application A <airport_reporting> as mapped to Application B <primary airport id>, while the exact values for 'Application A (1111) as mapped to application B (2259) will be contained in ORM Mapping Model (instance level).

The ORM Model (instance level) is used to establish precise cross-walk navigation between applications.

Also applicable at this instance level are local row specific mapping criteria and transform criteria.

It is important to note some of the background work in the area of semantic mapping that is now being promulgated as an ISO standard (ISO 20943-5 Semantic Metadata Mapping Procedure).

5 ISO Standard Work in Progress for Semantic Metadata Mapping Procedures

It is important to note that ISO IEC (WD) 20943-5 Semantic Metadata Mapping Procedure [3] attempts to formulate an orderly approach to setting up procedures for enabling metadata mappings based on the ISO 11179-3 Metadata Registries standard.

In brief, the ISO IEC (WD) 20943-5 Semantic Metadata Mapping Procedure defines types of Semantic heterogeneities [4] as: Hierarchical Difference - due to different level of detail, and generalization, specialization, composition, Decomposition; Domain Difference – due to different context and culture, and due to different context and culture; Lexical Difference – in different appearance, including synonyms, abbreviations, acronyms, case sensitivity, language, and variation; Syntactic Difference – due to different arrangement of parts, or ordering, delimiters, missing parts; and, Complicated Differences – due to different policies.

The ISO standard work is still in its beginning stages and needs to mature---particularly to take note of previously published work. This ISO Standard is expected to be published as a Technical Report in 2012.

6 Determining Equivalency – Probes

In a practical real implementation in the TOD project, a semantic mapping exercise was conducted for "Location", where a location is defined as "A spatial area

designated for purposes of transportation that is of regulatory interest to Transport Canada" e.g. Airport, Marine port, Rail station, Road Station etc.

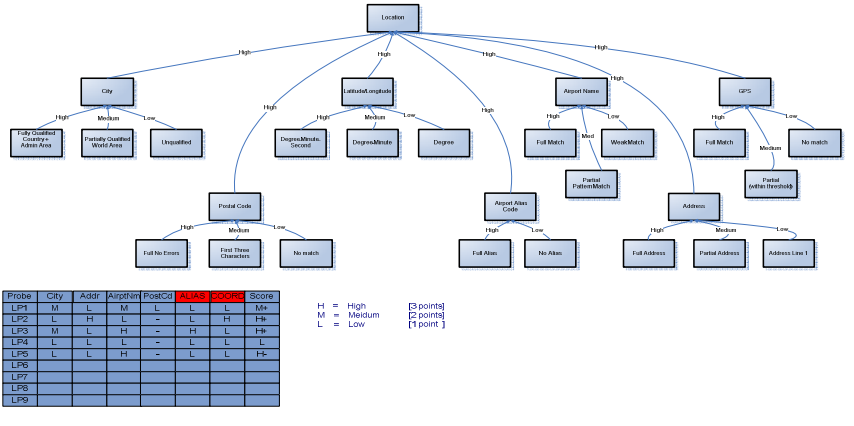
While it was a relatively simple task to establish metadata mappings for Location, the task of integrating the data at the value level proved to be a challenging task.

The objective was to establish value equivalency with each of the participating applications, each of which had their own identification scheme and alias code. A location alias was defined as: "Concordance or cross-reference of stated location with any other organization assigned identifiers or codes for airports, ports, train stations etc. e.g. IATA or ICAO codes for Airports, UN/LOCODE for trade including other organizations like US Federal Aviation Administration FAA etc.

While some locations had an IATA code, others had only ICAO codes, and if the locations were strictly local within Canada, they had the Canadian Flight Supplement (CFS) code etc. That means, a major airport say, Montreal, could have an IATA code, ICAO code and CFS code, while a smaller domestic only airport could only have a CFS code. The situation is exacerbated with the practice of the assignment of makeup alias codes by each of these organizations---conducted as an interim measure while a permanent code is in the approval process.

Here is a sample model that contributed to determining the harmonization of Location, using a decision tree approach, and, a set of probes that were written to investigate whether the given location was a member of a particular set or not. The "probes" were SQL code snippets in the form of User Defined Functions that would essentially "probe" the membership validity of a given instance in a desired set. See model along with probe scoring as in Fig. 4.

As mentioned earlier, not all locations had all aliases or geo coordinates available. The other identifying attributes like city name, address, airport name, postal code etc. were localized non-standard interpretations containing variants, which could not directly be used to establish equivalency. It was easy to establish equivalency on the alias codes and coordinates (again here within a threshold since each of the readings could be off by several decimal points---some coordinates may refer to the airport, some coordinates may refer to the city to which the airport belongs etc).



Probe	City	Addr	Airpfnm	PostCd	ALIAS	COORD	Score
LP1	M	L	M	L	L	L	M-
LP2	L	H	L	-	L	H	H-
LP3	M	L	H	-	H	L	H-
LP4	L	L	L	-	L	L	L
LP5	L	L	H	-	L	L	H-
LP6							
LP7							
LP8							
LP9							

Fig. 4. Probe Based Model for Determining Location Equivalency

In the example shown in Fig. 4, it was easy to establish equivalency confirmation using probe LP2 and LP3 provide a result of H+, which means we are able to confirm the equivalency of that location based on the 2 mandatory critical criteria matches ALIAS and COORD (geo-coordinates), while probe LP4 provided a no-match condition with an L rating since only low level matches were available for city, Address and Airport names—meaning there were syntactical variations. To standardize the syntactic variations, each of the columns being compared had their whitespaces, special characters, noise characters removed and the syntactic names compressed for matching. But these would not take care of misspelled or flipped characters e.g. 'abc' vs 'acb'.

We assigned a score of 3 for High, 2 for Medium, and 1 for Low, but doubled the number for the definitive attributes of Alias and Geocode matches. Then we added up the individual score from the probes and divided by the total max score of the attributes involved or where they were available to come up with a final score.

We found that we could trust the ratings at H- and over, which were put through the automated transfer cycle. Those probe results that produced anything less than H- had to go through a manual review. We also found many of the M+ ratings acceptable with minor corrections made to syntax string values.

A probe orchestration sequence model was established as is shown in the later section. This enabled conditional branches and probe activation based on the results (post conditions) of the previous probe result.

7 An Implementation of the ORM Mapping Models Probes

To automate much of the probe processing and Extract Transform and Loading processes, the ORM Mapping models shown earlier were implemented in ISO SQL 9075 Database Language SQL standard implementations or a variant thereof. The attribute based schema is shown in Fig. 5, which is an implementation of the ORM

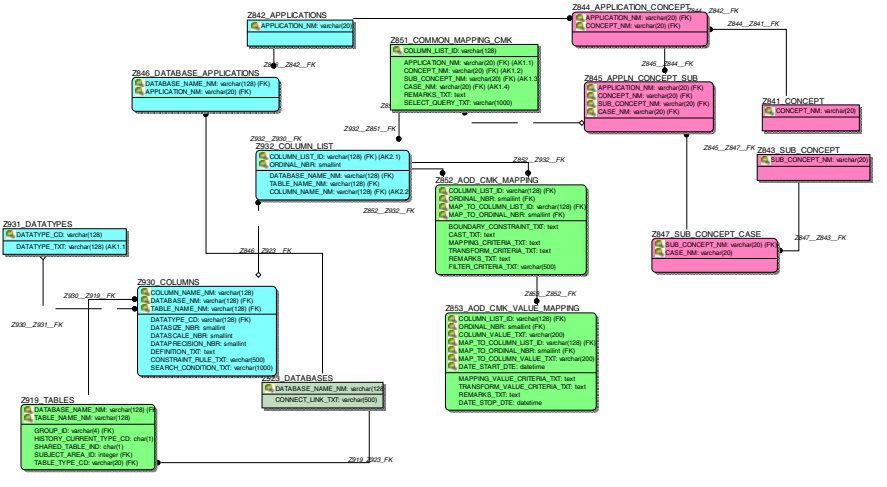


Fig. 5. Implementation Model of ORM Mapping Models (CMK)

mapping models that has proven itself over many projects through thick and thin over the last 10 years in Oracle and SQL Server DBMSs. Each concept like “Location”, “Carrier”, ‘Aircraft”, “Make-Model” has mappings established from various systems into the TOD, where TOD provides a central cross-reference and bridge to enable navigation across various participating applications via the Centralized mapping model, affectionately termed as the CMK (Common Mapping Key) Model.

8 Audit and Control – for Loading and Probe Tracking

Due to the many probe attempts involved, it is necessary to track the complete process of Extraction Transformation and Loading as the main thread, and, as a secondary thread, the tracking of the orchestrations of probe sequences. An implementation model of how such tracking was achieved is shown as in Fig. 6:

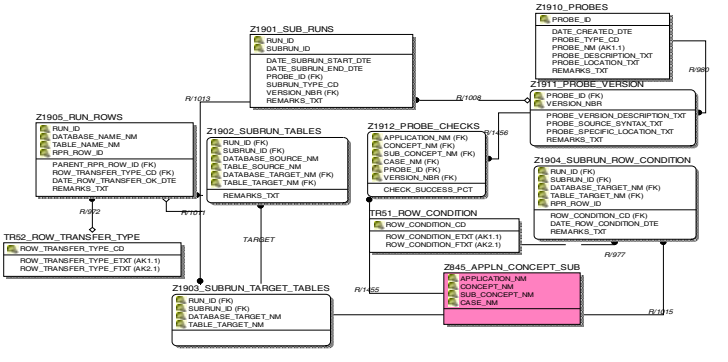


Fig. 6. Probe Tracking Model (partial)

A “run” is established for the Extraction, Transform and Loading of data being migrated from source to target based on the ORM Mapping Models. This ”run” follows the Audit and Control Schema parameters as established in [5] the Data Quality Firewall architecture for an Advanced Generation Data Warehouse.

Within this “run” there could be several probe “sub-runs”. The model in Fig. 6 tracks the results from each probe evaluating the post-conditions and rating whether that row under “probing” is ready for insertion as a new row occurrence of “Location” or is really an already established “location” occurrence.

Based on the probe score, the sub-run model would essentially be the decision-maker for the incoming row that is to be merged or mapped, or even whether to be accepted for integration. In other words, this part of the model essentially provides the intelligence for an SQL ‘MERGE’ or ‘INSERT’ operation that follows.

We did come across occasions where we needed to spice up some of the specialized versions of the probes after ‘learning’ from a probe run, and to incorporate this intelligence in the new version. That is why versioning played an important part of the probe sequence.

9 Use Case of a Successful Implementation in Practice

The use case as part of the TOD project is in the process of successfully analyzing data from multiple sources to build a Master data set consisting of "multiple versions" versus the common notion of a "single version". Why? We felt we could not ignore the local version that is in use in a particular application, and that there was no mandate for that local application to switch to a master data version (single version). Reality dictated that we keep track of each of the local versions, yet establish bridges across the multiple versions in order to establish navigation and access as shown below in Fig. 7. [9]:

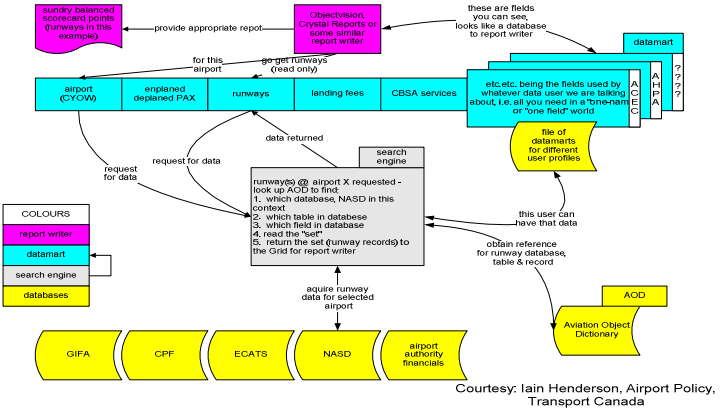


Fig. 7. Using TOD to Bridge Across Heterogeneous Applications

The application set essentially houses the metadata mapping intelligence in the TOD master database and the Common Mapping Key database. Wrappers are dynamically generated from the metadata in these databases for loading, or for remote retrievals as needed on demand. Business Intelligence tools are used to pull in and analyze the available data housed in heterogeneous databases at their own sites.

In summary, the TOD project at Transport Canada is successfully using Object Role Modelling [6] as their base semantic modeling method, interacting with the user community with natural language sentence constructs, and using ISO TR9007 [7] as the kingpin pivot for all rule declarations---at the same time, meeting the realistic corporate requirements of TC naming standards, attribute grouped data models using ERWin, and SQLServer and Oracle DBMSs based on maximizing essential ISO 9075 Database Language SQL [8] Schema declarative statements.

Acknowledgement

I would like to acknowledge and thank the Transport Canada TOD (Transport Object Dictionary) team, staff and users, Michel Villeneuve, Madhu Hundal, Tom Nash, Vasko Miovski, Dave McCutcheon, Rob Henkel, Wayne Shimoon, Jean-Pierre

Sabbagh El Rami, Jean Yves Cadieux, Iain Henderson, and Dave Dawson for their support and advice in establishing the necessary mapping algorithms and developing probes in support of this successful application.

References

1. Piprani, B.: Using ORM in an Ontology Based Approach for a Common Mapping Across Heterogeneous Applications. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM-WS 2007, Part I. LNCS, vol. 4805, pp. 647–656. Springer, Heidelberg (2007)
2. International Standard ISO IEC 11179:2003 Metadata Registries, International Standards Organization, Geneva
3. International Standard (WD) ISO IEC 20943-5:Semantic Metadata Mapping Procedure, International Standards Organization, Geneva
4. Semantic Metadata Mapping Procedure and Types of Semantic Heterogeneity, Tae-Sul-Seo, Korea Institute of Science and Technology Information, 12th Annual Forum for Metadata Registries, Seoul, Republic of Korea, June 18-19 (2009)
5. Meersman, R., Tari, Z., Herrero, P.: Using ORM-based models as a foundation for a data quality firewall in an advanced generation data warehouse. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4278, pp. 1148–1159. Springer, Heidelberg (2006)
6. Nijssen, G.M., Halpin, T.A.: Conceptual Schema and Relational Database Design. Prentice Hall, Victoria (1989)
7. van Griethuysen, J. (ed.): Technical Report on Concepts and Terminology for the Conceptual Schema and the Information Base. ISO Technical Report ISO IEC TR9007:1987. International Standards Organization, Geneva (1987)
8. International Standard ISO IEC 9075:1999. Database Language SQL. International Standards Organization, Geneva (1999)
9. Piprani, B.: Ontology Based Approach to a Common Mapping across Heterogeneous Systems. Presentation at Metadata Open Forum, New York (2007), http://metadataopenforum.org/index.php?id=2,0,0,1,0,0
10. Visser, J.: Finding nontrivial semantic matches between database schemas, Masters Thesis University of Twente Publications, Netherlands (July 2007), http://doc.utwente.nl/64138/
11. Garcia-Molina, H.: Entity resolution: Overview and challenges. In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.-W. (eds.) ER 2004. LNCS, vol. 3288, pp. 1–2. Springer, Heidelberg (2004)
12. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. IEEE Transaction on Knowledge and Data Engineering 19 (January 2007)
13. SERF, Stanford Entity Resolution Framework, http://infolab.stanford.edu/serf/
14. Xu, L., Embley, D.W.: Using Schema Mapping to Facilitate Data Integration, http://www.deg.byu.edu/papers/integration.ER03.pdf
15. Ram, S., Park, J.: Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflicts. IEEE Transactions on Knowledge and Data Engineering 16(2) (February 2004)