21st IEEE International Symposium on Computer-Based Medical Systems

Automatic Generation of Integration and Preprocessing Ontologies for Biomedical Sources in a Distributed Scenario

Alberto Anguita, David Pérez-Rey, José Crespo and Víctor Maojo
Biomedical Informatics Group, Artificial Intelligence Laboratory,
School of Computer Science, Universidad Politécnica de Madrid
Campus de Montegancedo, s/n. 28660 Boadilla del Monte, Madrid
{aanguita, dperez, jcrespo, vmaojo}@infomed.dia.fi.upm.es

Abstract

Access to a large number of remote data sources has boosted research in biomedicine, where different biological and clinical research projects are based on collaborative efforts among international organizations. In this scenario, the authors have developed various methods and tools in the area of database integration, using an ontological approach. This paper describes a method to automatically generate preprocessing structures (ontologies) within an ontology-based KDD model. These ontologies are obtained from the analysis of data sources, searching for: (i) valid numerical ranges (using clustering techniques), (ii) different scales, (iii) synonym transformations based on known dictionaries and (iv) typographical errors. To test the method, experiments were carried out with four biomedical databases —containing rheumatoid arthritis, gene expression patterns, biological processes and breast cancer patients— proving the performance of the approach. This method supports experts in data analysis processes, facilitating the detection of inconsistencies.

1. Introduction

Structural and semantic differences are among the main reasons for the complexity of managing data from heterogeneous sources, particularly in some specific domains such as biomedicine [1]. In this area, collaborative research among remote institutions has been responsible for completing the Human Genome Project before schedule. For these projects, researchers need to access and retrieve information from a large number of private and around 900 public databases. Such scenario presents different challenges for data integration and knowledge discovery in databases (KDD) research, particularly related to semantic heterogeneity

Although the scientific literature recognizes the relevance of other steps in the KDD [2], more efforts have been dedicated to the data mining process. Considering the preprocessing phase of the classical KDD methodology, its main goal is to provide data miners with 'clean' data —free of syntactic and semantic inconsistencies—, eliminating noise from the original dataset. In the highly heterogeneous and distributed scenario presented above, preprocessing is a key issue. Given the various physical and conceptual differences between biological and clinical data, this process has been usually performed manually by domain experts, since some specific background and expertise are needed.

We present in this paper a novel approach to automatically detect inconsistencies and store the corresponding transformations in a formal structure, i.e. an ontology [3]. This research is based on previous work carried out by the authors on ontology-based integration and preprocessing [4][5]. The objective of this current approach is to support KDD professionals by automatically generating instances of preprocessing and integration ontologies that can be used in the distributed KDD processes.

2. Background

Quality of data is a major concern in KDD processes. The presence of data inconsistencies might lead to inaccurate or useless knowledge in the discovery process. Reviewing the scientific literature—particularly related to distributed systems—, few efforts have been accomplished regarding cleaning of data inconsistencies and quality assurance. When dealing with the integration of data from distributed and heterogeneous systems, differences in formats and patterns are serious problems that should be addressed by KDD researchers. ETL (Extract Transform and Load) modules are the most common tools for

1063-7125/08 $25.00 © 2008 IEEE
DOI 10.1109/CBMS.2008.71

preprocessing. However they are intended for centralized environments (Data Warehouses), using proprietary APIs (Application Programming Interfaces) and very specific for a certain domain–e.g. standardizing addresses. So they cannot be used in distributed frameworks where there exist few preprocessing researches [6][7][8].

In the context of heterogeneous and remote database scenarios, ontologies have been applied during the last years to develop new approaches for semantic-based data integration and preprocessing. Ontologies facilitate researchers the description of a shared domain using a formal foundation, providing an intuitive framework and permitting easy data sharing among different sources. As regards, the use of ontologies has been analyzed to enhance the overall performance of the different KDD process [9]. Table 1 lists some examples of ontology-based systems used for data integration and preprocessing. Two categories are considered, according to the granularity of the approach: schema and instance level.

Table 1. List of ontology-based systems in integration and preprocessing

KDD phase	Systems
Integration at a schema level	[10], D2RMAP, SEMEDA, KAON, ONTOFUSION [5]
Preprocessing and Integration at an instance level	[11], [12], ONTOCLEAN, ONTODATACLEAN [4]

The majority of these systems follow an approach with two sequential steps: (i) Detection and (ii) Resolution. Although the resolution phase has been successfully automated in these systems, the detection phase is still carried out manually. It is a very time consuming task, and it is error prone. In fact, manual inconsistency detection is unapproachable in huge biomedical resources such as the Unified Medical Language System (UMLS) [13]—a comprehensive biomedical vocabulary and nomenclature containing more than 1.3 million concepts and 6.4 million unique concept names in the last 2007AA release.

Automatic methods for inconsistency detection at schema and instance level are necessary, but few efforts can be found in the literature [14][15]. Authors of [16] developed a framework cleaning retrieved from remote Internet sources, based on the idea of detection-resolution. Other works, such as [17], adopt the same two-module design, but simply provide improved data clustering algorithms to increase the quality of the detection module. One aspect that all these systems lack is intuitiveness together with easy sharing of results with other tools. Our framework for data cleaning and inconsistency detection, based on ontologies, provides both aspects. This feature is an important advantage if we take into account that the expected users of this application are experts on the data being preprocessed, but not necessarily experts on APIs or data formats (e.g. XML).

3. Distributed and Ontology-Based KDD Approach

This paper is the extension of two ontology-based systems previously developed by the authors to address the integration and preprocessing of distributed data: (i) OntoFusion [5], a system developed to carry out integration at the schema level and (ii) OntoDataClean [4], a system carried out to perform integration and preprocessing at the instance-level. They were developed and used by the authors in the context of two projects funded by the European Commission, the INFOBIOMED Network of Excellence and the ACGT project —Advanced Clinico-Genomic Trials on Cancer.

The previously mentioned project carried out by the authors, OntoFusion was designed to address semantic heterogeneity among data sources. To achieve information integration, similarities must be recognized and presented to the users. In OntoFusion, virtual schemas of databases are represented as ontologies, i.e. every concept in a physical database is mapped to a concept of a specific domain ontology [18].

OntoDataClean was developed to cover preprocessing and integration at instance-level tasks within OntoFusion. Following a similar ontology-based and “virtual” approach, OntoDataClean also uses ontologies as structures to store the information needed to carry out data transformations. Using these metadata, OntoDataClean can cope with six types of inconsistencies: (i) Missing values, (ii) Format, (iii) Scale, (iv) Pattern, (v) Synonyms and (vi) Duplicates. After the incorporation of these preprocessing ontologies, the resolution mechanism performs the required data transformations automatically each time a query is launched to the system.

Figure 1 presents the situation of OntoDataClean within a distributed and ontology-based KDD methodology. The distributed approach of sources is maintained until the data mining phase using ontologies as formal support. Although experts detecting inconsistencies in this framework may use ontology editors such as Protégé, SWOOP, KAON2 and others, specific tools to partially automate this process are needed.

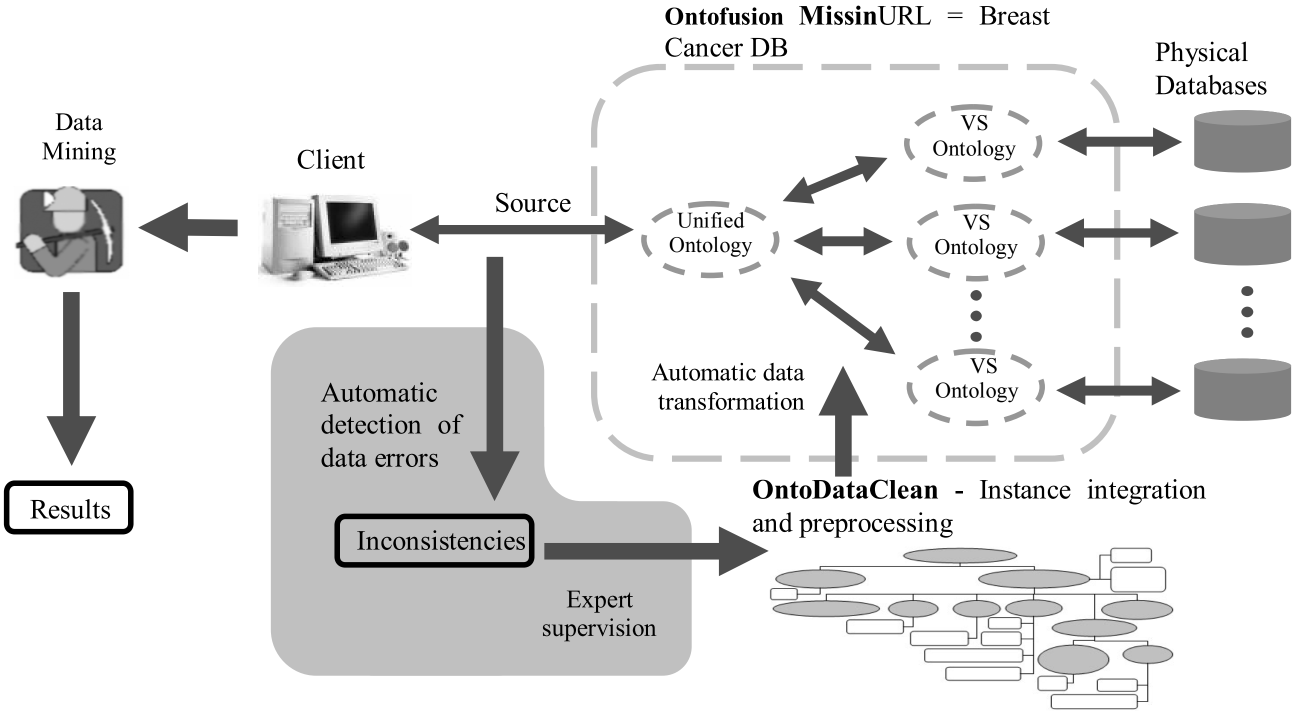
Ontofusion MissinURL = Breast Cancer DB
Physical Databases
Data Mining
Client
Source
Unified Ontology
VS Ontology
VS Ontology
VS Ontology
Automatic data transformation
Automatic detection of data errors
Results
Inconsistencies
OntoDataClean - Instance integration and preprocessing
Expert supervision

Figure 1. Distributed and ontology-based KDD approach (in grey the semi-automatic generation of preprocessing ontologies)

4. Generation of Integration and Preprocessing Ontologies

Following the approach stated in Figure 1, OntoDataClean is composed of a detection module —where this work is centered— and a resolution module —described in section 3. The detection module is devoted to: (i) analyzing data sources, (ii) identifying the corresponding inconsistencies and (iii) automatically generating ontology instances that store the information about data transformations. These instances will then be at user disposal to review the generated cleaning model and, in case it is required, adjust it. The use of ontologies facilitates this stage as it offers users with a deep yet intuitive view of the cleaning model.
The tool employs four different algorithms for inconsistency detection and subsequent preprocessing ontology generation. In every case, a factor of correctness is also calculated, so user can better evaluate the adequacy of results. Next subsections briefly describe such algorithms and the ontologies that they generate.

4.1 Detection of valid numerical ranges

Numerical values tend to lie within a statistically more probable range—e.g. age ranges between 0 and around 100. The algorithm uses the Mahalanobis distance to evaluate which values are to be considered outliers [19], since, unlikely the Euclidean distance, it does take into account the dispersion of the values. Figure 2 shows the formula to calculate the Mahalanobis distance, given a point x.

$$D_M(x)=\sqrt{(x-\mu)\sum^{-1}(x-\mu)}$$

Figure 2. Calculation of the Mahalanobis distance

The generated preprocessing ontology in this case is composed of a MissingValue class. It eliminates rows where outlier values appear. The factor of confidence in this case is calculated comparing the number of erased values to the total number of values of the initial set.

4.2 Detection of fields requiring typographical corrections

A database field containing string values is always predisposed to suffer of typographical errors. The detection module analyzes textual fields in search of values that require a typographical correction. The hypothesis is that these values differ little from their corresponding correct value, and their proportion compared to the correct value is lower than a given threshold. Following this premise, the tool analyses textual fields and produces a list of candidates to typographical correction.

An ontology with instances of Value class for each of the mentioned candidates is created. It substitutes the value suspect of being wrong with the supposedly correct value. The factor of confidence for this algorithm is given by the ratio of the number of appearances of a misspelled term against the number of appearances of its correct form.

4.3 Detection of fields requiring a dictionary

Many terms, and especially in the biomedical domain, have one or several synonyms, usually being just one of them the preferred one. This is reflected in specific purpose dictionaries containing large collections of terms and synonyms, such as UMLS.

Proper data integration requires syntactically equivalent values. In order to achieve this, synonym heterogeneities must be solved by translating synonyms into the preferred values. Our approach analyses fields with textual values and searches the terms that appear in the UMLS, but are not the preferred terms. If this is the case, an ontology instance, consisting on a single instance of the class SynonymDatabase, transforms these terms by means of the mentioned dictionary. The factor of confidence depends on the number of non-preferred terms compared to preferred-terms. The lower this ratio is, the greater the factor of confidence becomes.

4.4 Scale Detection

OntoDataClean is able to find scale heterogeneity among a set of databases by comparing specific statistical markers—namely, mean and variance. If divergences are found, it calculates linear algebraic transformations to solve them—homogenize the previous values.

The result is an ontology instance specifying a scale transformation. The deducted linear algebraic transformation will be the only attribute of this instance. After applying this method, a test that analyzes the similarity of the deciles, the kurtosis coefficients and the symmetry coefficients of the sets is performed to evaluate their degree of similarity.

5. Results

The following paragraphs illustrate two cases from the Surveillance Epidemiology and End Results (SEER) database from the National Cancer Institute. Data is shown prior and after transformation by the automatically generated preprocessing ontologies. Pictures describing such ontologies are also shown.

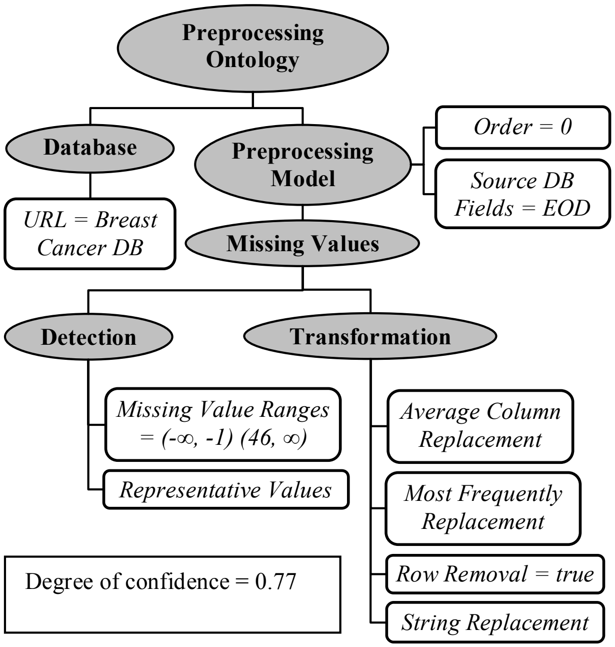

Figure 3. Automatically generated preprocessing ontology to erase missing values

In the first example, a database containing values representing tumor sizes is analyzed. These values generally range from 0 to approximately 50, but the values 997, 998 and 999 are employed to store special values—diffusion, not detected, etc. Although they are not errors, these special values should be erased in order to perform a correct integration with other databases containing semantically equivalent data. As it can be seen in figure 3 the generated preprocessing ontology suggests a missing value transformation that erases the mentioned values, considering that only values up to 45 can be considered as valid. Figure 4 shows the distributions of tumor size values before and after applying the automatically generated preprocessing ontology.

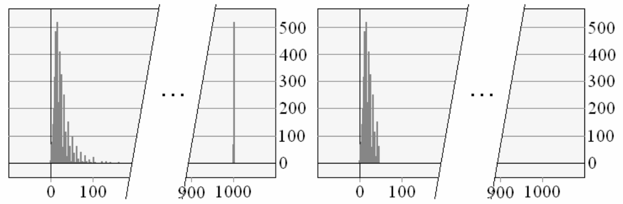

Figure 4. 'Tumor Size' value distribution of the "Breast cancer database". The first graph represents the original data and the second the data queries through OntoDataClean using the generated ontology

The second case consisted on the homogenization of the numeric values of two databases. Previous analysis of data concluded that the two sources contained semantically equivalent data. They both expressed the year of birth of patients, but whereas the first source contained this actual value, the second source stored integers representing the year of birth counting from 1800. In order to allow proper integration of data, syntactic dissimilarities must be eliminated. The tool correctly detected this inconsistency and generated an ontology that modified the values of the second source by means of applying a linear algebraic expression.

Figure 5 depicts the generated preprocessing ontology, which acts on the values from DB2 in order to make them syntactically homogeneous with the values from DB1.

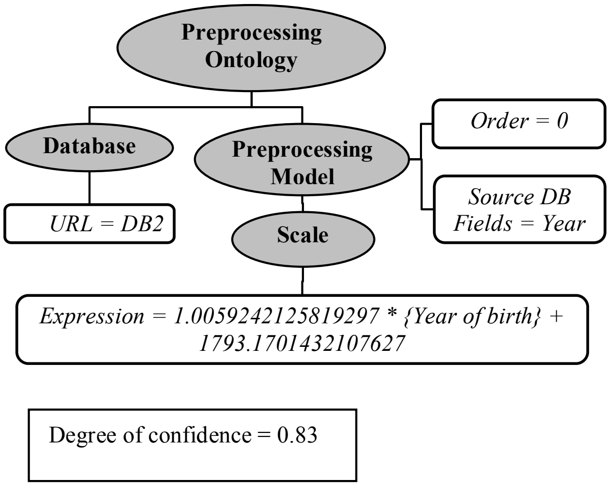
Preprocessing Ontology
Database
Preprocessing Model
Order = 0
Source DB Fields = Year
URL = DB2
Scale
Expression = 1.0059242125819297 * {Year of birth} + 1793.1701432107627
Degree of confidence = 0.83

Figure 5. Automatically generated preprocessing ontology to integrate DB 1 and DB 2

Figure 6 shows the values from DB1 and DB2 before and after using the mentioned preprocessing ontology. On the left side the original values from both databases are shown—DB1 on top and DB2 on bottom. On the right side, values are shown again after proper transformation has been applied by the tool using the previous preprocessing ontology.

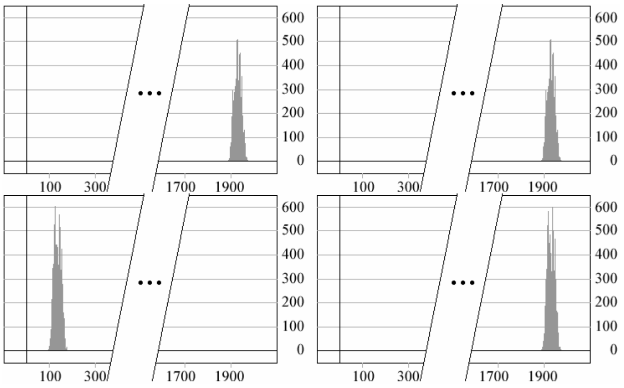

Figure 6. Syntactic dissimilarity in the Year of Birth due to differences in scale (left) is resolved by the preprocessing ontology (right)

These experiments showed the correct behavior of the tool using real examples. This suggests the viability of automating the inconsistency detection phase using an ontological approach. It must be noted that the automatically generated preprocessing ontologies must be always supervised by a domain expert, and may sometimes require of adjustments for proper data manipulation. Nevertheless, results provide valuable and intuitive suggestions for data preprocessing and integration.

6. Conclusions and Future Work

In this paper a new method specially suited for biomedical sources to detect inconsistencies and generate preprocessing ontologies has been presented. This technique is embedded within a global ontology-based system used for KDD. Due to the special characteristics of biomedical information systems, a distributed approach has been adopted for the KDD model.

The automatic generation of preprocessing ontologies (section 4) is used to analyze ranges, scales, synonyms and typographical errors in data sources. The output is an ontology containing suggestions about transformations together with a factor of confidence. After their revision by a data expert, they are included in a global KDD model (section 3) to complete the detection phase. Preprocessing ontologies together with the corresponding virtual schemas facilitate information retrieval, transforming the data when users query the heterogeneous sources through a common interface. Two types of data sets from a biomedical database were successfully preprocessed using our system. The resulting ontologies were incorporated into the global model, enhancing the information retrieval.

Compared to other works, such as the ones presented in [12] and [15], our system provided a broader range of preprocessing suggestions. The generated preprocessing ontologies provide an open an intuitive approach for data sharing and management. Eliminating proprietary formats allows easier interoperability among different institutions and researchers, facilitating the collaborative work.

We are currently following this work in various directions. The first one is the automatic generation of mappings for the schema-level integration. This investigation aims to provide a complete support system for ontology-based integration and preprocessing. GRID computing environments are also being considered in the context of the ACGT project. Another option is to modify the preprocessing ontologies, according to the data mining algorithm to be used. Such approach might offer the possibility of transforming the data according to specific machine learning model.

7. Acknowledgements

This work was funded by the Ministry of Education, Spain (OntoMineBase project, TSI2006-13021-C02-01) and the European Commission **(**the INFOBIOMED NoE (FP6-2002-IST-507585) and the ACGT integrated project (FP6-2005-IST-026996)).

8. References

[1] D. Gurwitz, J.E. Lunshof, and R.B. Altman, "A call for the creation of personalized medicine databases", *Nature Reviews Drug Discovery*, *5* **1,** 2006, pp. 23-26.

[2] T. Dasu, and T. Jonson, *Exploratory Data Mining and Data Cleaning*. John Wiley, 2003.

[3] T.R. Gruber, "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition, 5* **2**, 1993, pp. 199-220.

[4] D. Perez-Rey, A. Anguita, and J. Crespo, "OntoDataClean: Ontology-based Integration and Preprocessing of Distributed Data", *Lec. notes in Computer Science* **4345**, 2006, pp. 262-272.

[5] D. Perez-Rey, V. Maojo, M. Garcia-Remesal, R. Alonso-Calvo, H. Billhardt, F. Martin-Sanchez, and A. Sousa, "ONTOFUSION: Ontology-Based Integration of Genomic and Clinical Databases", *Computers in Biology and Medicine* **36**, 2006, pp. 712-730.

[6] E. Rahm, and H.H. Do, "Data cleaning: problems and current approaches", *IEEE Bulletin of the Technical Committee on Data Engineering, 23* **4**, 2001, pp. 3-13.

[7] W. Sujansky, "Heterogeneous Database Integration in Biomedicine", *Journal of Biomedical Informatics, 34* **4**, 2001, pp. 285-298.

[8] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, "Ontology-based integration of information – A survey of existing approaches", In *Workshop on Ontologies and Information Sharing (IJCAI-01)*, 2001, pp. 108-117.

[9] H. Cespivova, J. Rauch, V. Svatek, M. Kejkula, and M. Tomeckova, "Roles of Medical Ontology in Association Mining CRISP-DM Cycle", In *Proceedings of the ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04),* Pisa, 2004.

[10] A. Silvescu, J. Reinoso-Castillo, and V. Honavar, "Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed, Autonomous Data Sources", *Proceedings of the IJCAI*, 2001.

[11] J. Phillips, and B.G. Buchanan, "Ontology-guided knowledge discovery in databases", *Proceedings of the International Conf. Knowledge Capture,* Canada, 2001.

[12] Z. Kedad, and E. Métais, "Ontology-based Data Cleaning", *Lec. notes in Computer Science* **2553,** 2002, pp. 137-149.

[13] C. Lindberg, "The Unified Medical Language System (UMLS) of the National Library of Medicine", *Journal of the American Medical Record Association, 61* **5**, 1990, pp. 40-42.

[14] E. Rahm, and P.A. Bernstein, "A survey of approaches to automatic schema matching", *VLDB Journal, 10* **4**, 2001, pp. 334–350.

[15] J.I. Maletic, and A. Marcus, "Data cleansing: Beyond integrity analysis", In *Proceedings of the 5th Conference on Information Quality*, 2000, pp. 200–209.

[16] Y. Lu, and H. Jiang, "A Framework for Efficient Inconsistency Detection in a Grid and Internet-Scale Environment", In *Proceedings of the 14th IEEE International Symposium on High Performance Distributed Computing (HPDC-14)*, 2005, pp. 318-319.

[17] P. Anokhin, and A. Motro, "Data integration: Inconsistency detection and resolution based on source properties", In *Proceedings FMII-01, International Workshop on Foundations of Models for Information Integration*, 2001.

[18] H. Billhardt, J. Crespo, V. Maojo, F. Martín, and J.L. Maté, "A New Method for Unifying Heterogeneous Databases", *Proceedings of the ISMDA*, 2001, pp. 54-61.

[19] R. De Maesschalck, D. Jouan-Rimbaud, and D.L. Massart, "The Mahalanobis distance", *Chemometrics and Intelligent Laboratory Systems* **50**, 2002, pp. 1-18.