Computer Science and Information Systems 12(1):135–160 DOI: 10.2298/CSIS140325087D

Academic Data Warehouse Design Using a Hybrid Methodology

Francesco Di Tria, Ezio Lefons, and Filippo Tangorra

Dipartimento di Informatica
Università degli Studi di Bari Aldo Moro
Via Orabona 4, 70125 Bari - ITALY
{francesco.ditria, ezio.lefons, filippo.tangorra}@uniba.it

Abstract. In the last years, data warehousing has got attention from Universities which are now adopting business intelligence solutions in order to analyze crucial aspects of the academic context. In this paper, we present the architecture of a Business Intelligence system for academic organizations. Then, we illustrate the design process of the data warehouse devoted to the analysis of the main factors affecting the importance and the quality level of every University, such as the evaluation of the Research and the Didactics. The design process we describe is based on a hybrid methodology that is largely automatic and relies on an ontological approach for the integration of the different data sources.

Keywords: Business intelligence system; decision making; data warehouse architecture; multidimensional modelling.

1. Introduction

In the last years, also Universities have accepted to adopt Business Intelligence systems [1, 2, 3, 4, 5, 6] and to develop data warehouses devoted to produce significant information to be used in their strategic decision making. The aim is to improve business processes of the academic information system, also using web-based environments [7]. Nonetheless, these organizations have specific purposes that differ considerably from those of enterprises and companies [8]. Indeed, typical objectives affecting the management of a University are: offering a better quality of the instruction; managing employees and human resources; managing economic-financial institutions; avoiding wastes, and increasing scientific publications and research projects.

Given these business goals, university decision makers are always interested in the possibility to timely make the best business decisions on the basis of historical data available in a unique and updated source of information. The main problem to be faced in the realization of the Academic Business Intelligence system is that each University has own legacy databases of historical data. Moreover, these databases are independent from each other, producing data redundancy and inconsistency. For example, the list of the professors is included into the database inherent the didactics offer. The same list is repeated in the database related to the university staff. It follows the need of data integration in order to provide useful information for analytic purposes. Also our

University meets the previously described conditions. Therefore, we decided to develop a data warehouse integrating all the present and historical data sources. The data warehouse covers different departmental areas and, therefore, is composed of several data marts. The most important topics of analysis regard (a) the quality of the Didactics, which aims at evaluating the performance of the students and at detecting the most productive courses of degree, and (b) the state of the Research, which aims at evaluating the scientific production of departments [9].

Because of the high complexity of the design process, which takes into account several data sources, and the necessity to effectively map business goals against available data sources, traditional data warehouse design methodologies do not suffice [10]. In fact, traditional methodologies are based on two opposite approaches. The one is data-oriented and aims to realize the data warehouse mainly through a reengineering process of the well-structured data sources solely, while minimizing the involvement of end users. The other is requirement-oriented and aims to realize the data warehouse only on the basis of business goals expressed by end users, with no regard to the information obtainable from data sources. Indeed, the requirement-driven methodologies may lead to a data warehouse conceptual schema inconsistent with data sources; on the other hand, data-driven methodologies may discard interesting user requirements [11]. Moreover, automation in design process helps designers in avoiding repetitive tasks, especially when a new data source is added [12].

For these reasons, we adopted a hybrid methodology that allows to define multidimensional schemas by first considering user requirements and, then, reconciling them against data sources. The core of the methodology is a multidimensional model providing a graph-oriented representation of the relational schema obtained through an ontology-based integration of different data sources. The steps of the design process are largely automatic, for they rely on a set of constraints derived from the requirement analysis. Such constraints allow performing a reengineering of the integrated source schema in a supervised way. At the end, the resulting schema is validated on the basis of a preliminary workload.

In this paper, we present the Research and the Didactics data marts realized on the basis of the hybrid methodology, in order to perform analysis of both the scientific publications and the students' performance in our academic environment. The Research data mart covers the needs related to the evaluation of the results gained by university staff involved in the Research activity, while the Didactics data mart is mainly devoted to the extraction of information about the career of the students in our Athenæum.

The paper is structured as follows. Section 2 presents related work about academic data warehouses. Section 3 shows the architecture of our academic Business Intelligence system. Section 4 illustrates the design of the academic data warehouse and, in particular, the Research data mart, for it provides the discussion about schema integration problems. Section 5 shows the Business Intelligence applications developed for evaluating the university activities. At last, Section 6 contains some our concluding remarks.

2. Related work

In [13], the authors propose the term education data warehouse (EDW) to describe a health system devoted to support the creation of individualized learning paths and to allow analyses about the career of physicians over time. The data warehouse is mainly used to collect and integrate data coming from different training programs and from national electronic databases. Such integration is encouraged by emerging standards for health professions, since these standards are lowering the barriers among medical institutional organizations [14].

The trend of adopting data warehouses for academic health systems in confirmed in [15], where the design experience in the University of Michigan Health System is reported. Here, the data warehouse is obtained through the integration of clinical and financial data, in order to understand the financial implications of clinical decisions in the care of patients. The underlying assumption is that clinicians may take better decisions when they know the costs of a particular practice and can identify alternative practices.

A similar case is that of the University of Virginia Health System, where the data warehouse is used to provide clinicians and researchers with direct and rapid access to retrospective clinical and administrative patient data [16]. In addition, they use the data warehouse also for educational and research aims, as it serves to face informatics issues —such as data capture—and to perform exploratory analyses of healthcare problems.

An interesting system devoted to the evaluation of performance in academic environments is presented in [17], where the authors illustrate a decision support system to carry out statistical analyses about the performance of students at course, program, department, school, and university levels. Here, the novelty is the possibility to use the collected data—which offer an in-depth view of factors affecting performance in universities—to define new academic performance evaluation criteria.

3. Architecture of the academic system

In general Business Intelligence consists of methodologies and technologies that support companies to obtain information about their own business processes. A Business Intelligence System in the University context aims in particular to understand the students' performance, the teaching staff's productivity, and the Research and Didactics quality. So, a University data warehouse represents a unique system of analysis available to the supervisory staff of the Athenæum and to single organizational and administrative structures, such as departments and secretariats for the students. Moreover, such a system is also able to supply in real time data to external information agencies devoted to control the results reached by the University.

Figure 1 shows the system architecture we developed for analytic purposes in our University. It consists of four levels: (1) the source relational databases, (2) the tool to extract data from source in order to feed the data warehouse, (3) the data warehouse divided into independent data marts, and (4) the OLAP (On Line Analytical Processing) layer which includes the applications to be used by decision makers for developing reports.

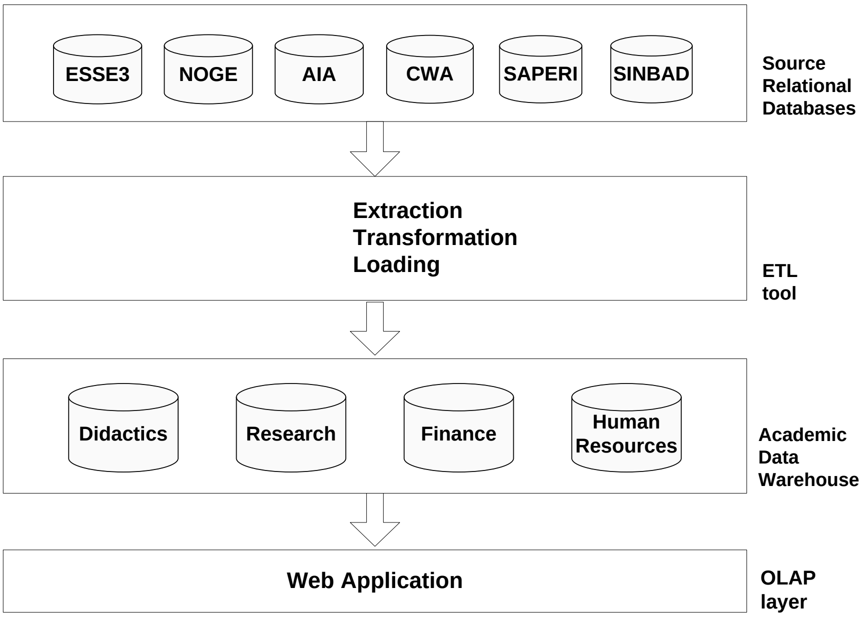


Figure 1. Academic Business Intelligence System

The source databases, which contain transactional data, are:

- ESSE3 (Secretary and Services for Students) is the new database that supports all the didactic curricula, and administrative processes and services to the students with the respect of the didactic autonomy of the University.
- NOGE (NOt ManaGEd) is a secondary old database that stores residual historical data about students enrolled before the ESSE3 introduction.
- AIA (Athenæum Integrated Accounting) is the integrated financial management system that considers the University as a business company that distributes specialized services (Research and Didactics, for example).
- CWA (Careers and Wages of Athenæum) takes care of the legal and economic management of the university personnel.
- SAPERI is the database of the scientific research competence of the University. It also includes publications and patents of researchers. These data concur, among other things, to construct the athenæum yearbook.
- SINBAD is the system for the management of the athenæum research projects.

The data warehouse is composed of a set of data marts to model the following academic departmental areas:

- *Didactics.* This data mart contains data about the career of the students of the Athenæum. Moreover, there is information on the University formation offer structured in Degree Courses.
- *Research.* The data mart contains awarded research projects and applications for research grants. It also contains data on components and location of every research project.

- *Finance.* This data mart is devoted to run twofold analyses: (a) the analysis of financial documents, and (b) the analysis of general and analytic economic movements.
- *Human Resource.* The model produced for this area allows investigating the legal-economic careers and wages of the academic personnel. Moreover, it allows extracting information related to the functions, activities, and location of the academic, administrative, and technical personnel.

4. Academic data warehouse

In this section, we illustrate the design process of the academic data warehouse. In particular, the case study explains the design of the Research data mart able to support the evaluation of the research activity in our University.

The University of Bari is equipped with an internal team for the evaluation of the Administration, the Didactics, and the Research in order to verify the correct utilization of the public resources. The team needs to gather and examine data for the evaluation of the several didactic and research activities held in the University. It periodically carries out technical reports to be transmitted to the national committee, which establishes the program guidelines and the quality goals to be satisfied. As an example, the team predisposes documentation about the state of the university education, the compliance with the rights to study, and the assurance of access to courses of studies.

Because of the complexity of the data warehouse to realize, we used a hybrid design methodology that aims to produce fact schemas, by considering at the same time both the user needs and the data sources. In this way, it is possible to obtain a data warehouse that is consistent with the available data without missing business goals. Moreover, hybrid methodologies usually provide also algorithms to define conceptual schemas in automatic way, in order to reduce implementation time, design errors, and wasting of time. We adopted our hybrid design methodology or the Graph-oriented Hybrid Multidimensional Model (GrHyMM, for short) [18, 19] whose phases are depicted in Figure 2 and described in the next subsections.

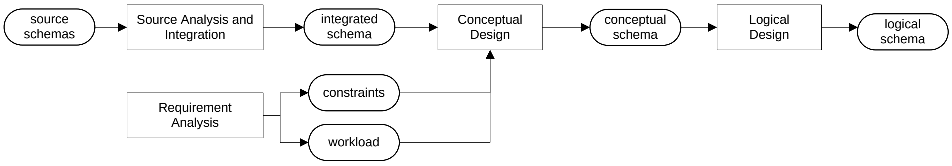


Figure 2. GrHyMM design methodology

4.1. Requirement analysis

The main aim of the decision makers is to evaluate the activities of the university researchers. To do so, they are mainly interested in knowing the number of scientific publications produced in the Departments. In this context, the fact of interest is the publication. This fact is an event that occurs each time one or more authors of the University receive the approval for the publication of a paper.

Decision makers are also interested in evaluating the quality of the Didactics. To this end, they need to analyze the students' curricula that include examinations and degrees. As a complementary analysis, they are interested in knowing the most and the least populated courses, in order to increase the number of students enrolled to the University, and the total amount earned from students' tax, in order to detect the least productive degree courses.

The case study needs the formalization of user requirements according to the *i** methodology [20]. First, possible actors must be identified. In our case, the Rector represents the maximum institutional figure in the University who decides on the university activities. The Rector is a member of the national council that (a) centralizes its own evaluation activity on Didactics and Research areas; and (b) develops and proposes methodologies and evaluation criteria for athenæums, and degree courses, finalized to the improvement of the quality of the Italian University system. The second actor is the Data warehouse that can be thought as an agent that aims to collect data from operational sources and to provide useful information to decision makers. For each decision maker, strategic goals must be identified. In the University context, the Rector aims at improving the Didactics and Research quality. This goal is modelled as a goal dependency from Rector to Data warehouse.

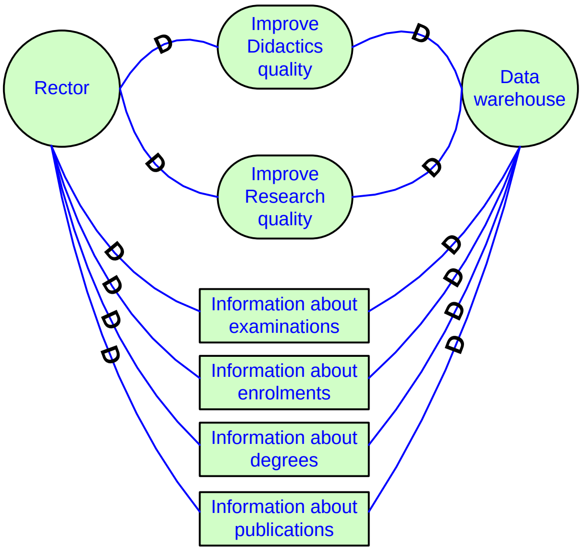

Rector
Improve Didactics quality
Data warehouse
Improve Research quality
Information about examinations
Information about enrolments
Information about degrees
Information about publications
D

Figure 3. Strategic dependency model

In order to accomplish the goals, resources are needed. The Rector needs to obtain information about publications, examinations, enrolments, and degrees. These requirements are modelled as resource dependencies from Rector to Data warehouse.

Now, the *i** strategic dependency model can be depicted with goal and resource dependencies (*cf.*, Figure 3).

Using this model, we next define the strategic rationale model for each actor except for the Data warehouse. As concerns the Rector, we have what follows. From the strategic goal "improve Research quality", decision goals are derived using a top-down approach to answer how strategic goals can be satisfied. On the turn, from decision goals, information goals must be derived using a top-down approach to define which information is needed for decision making. At last, the Rector accomplishes a set of tasks in order to achieve information goals, such as "to analyze number of publications per Department and scientific sector". These analyses represent information requirements. Part of the *i** strategic rationale model for the Rector is shown in Figure 4.

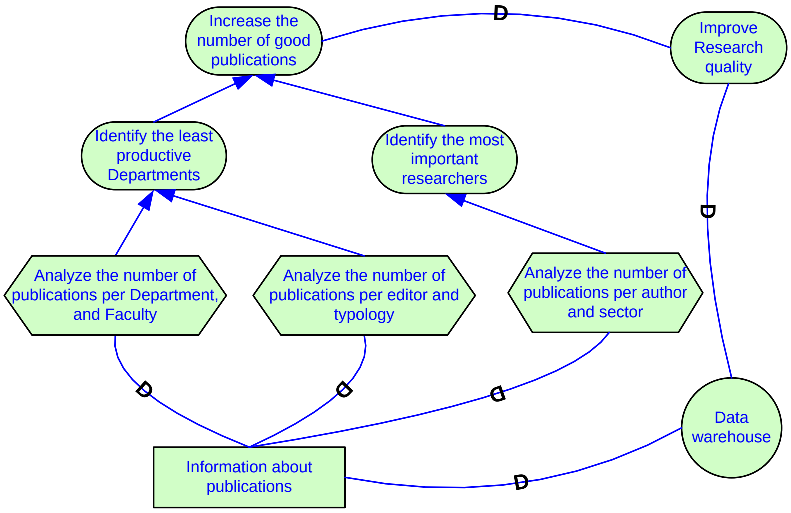

Increase the number of good publications
Improve Research quality
Identify the least productive Departments
Identify the most important researchers
Analyze the number of publications per Department, and Faculty
Analyze the number of publications per editor and typology
Analyze the number of publications per author and sector
Data warehouse
Information about publications
D
D
D
D
D
D

Figure 4. Part of the strategic rationale model for the Rector

So, using information requirements, we can define a *workload*, containing the typical analytical queries the decision makers intend to do. Of course, all the analyses should be done per year. Indeed, the most important points of view to analyze publications are the year of publication, author's name along with his/her affiliation, typology, and editor. In order to obtain reliable statistical data, the decision makers must be allowed to analyze publications related to the last ten years. Therefore, ten years represent the historicity level of these analyses. Then, part of the *workload* for the Research evaluation is:

- count of publications per author,
- count of publications per author and per typology in a given period of time,
- count of publications per Department, and
- count of publications per Sector.

Then, the strategic rationale model is created also for the Data warehouse actor. For each resource dependency, the Data warehouse must provide adequate information. In detail, it must have measures that are the resources used to provide the information required by decision makers (in this case, the Rector). Moreover, for each goal, a context of analysis must be provided by a task of the Data warehouse.

In Figure 5, we represent the goal of the Data warehouse in reference to the “information about publications” resource dependency: this goal is “provide information about publications” and it has no measures. In order to achieve the goal, it must execute the task “collect data about publications”, whereas its context of analysis must consider different resources such as author, Department, and so on.

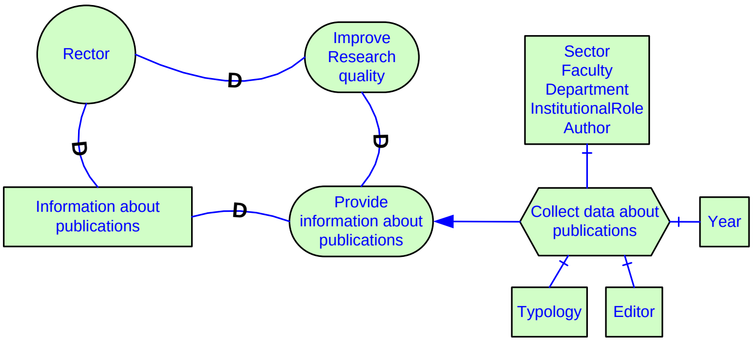

Rector
D
Improve Research quality
D
D
Sector
Faculty
Department
InstitutionalRole
Author
Information about publications
D
Provide information about publications
Collect data about publications
Year
Typology
Editor

Figure 5. Part of the strategic rationale model for the Data Warehouse

4.2. Source analysis and integration

The schemas of the different data sources must be analyzed and then reconciled, in order to obtain a global conceptual schema. The global conceptual schema resulting from the integration process must be then transformed into a relational schema, which constitutes the input to the next Conceptual Design. The integration strategy is based on an ontological approach [21] and, therefore, we need to produce an ontology for representing in formal way the main concepts of the domain of interest, along with their relationships. The ontology derived from OpenCyc [23] is shown in Figure 6.

As concerns the Research, the source databases are CWA and SAPERI. CWA aims at managing legal and economic data of the university personnel, while SAPERI stores data about scientific publications. As to the CWA database, we are interested in knowing, for each person, the full name, the institutional role in the University (which can be, for example, assistant professor, associate professor, full professor, …), the Sector representing the scientific area of membership, and the Department of affiliation. From the SAPERI database, we want to know the list of scientific publications, and, for each of them, the title, a brief description, its typology (that is, book chapter, journal paper, …), the publication year, the author(s), the language, the editor, the pages, the ISBN and DOI codes. Since the integration process relies on an ontological approach and, then, works at the conceptual level, we first need to represent the source databases according to the Entity/Relationship model (see Figure 7 for the essential schemas).

Now, we use the ontology as a vocabulary for the creation of definitions of each concept in the databases as reported in Table 1. Based on these definitions, we apply the similarity degree metrics [24] in order to automatically check whether entities refer to the same ontological concept. We assume that two entities represent the same concept if the threshold value 0.7 is gained. Given the lists L_1 and L_2 containing the definitions of two entities, the similarity degree d is given by

$$d(l,n,m) = 0.5 \times \frac{l+1}{l+n+2} + 0.5 \times \frac{l+1}{l+m+2}$$

where:

- l is the number of ontological equivalences $p \leftrightarrow q$ between predicates $p \in L_1$ and $q \in L_2$,
- n is the number of non ontologically-equivalent predicates p ($p \not\leftrightarrow q$) for some $p \in L_1$ and for all $q \in L_2$, (in such cases, $p \notin L_2$ for $p \leftrightarrow q$ holds true if $p=q$), and
- m is the number of non ontologically-equivalent predicates q ($p \not\leftrightarrow q$) for some $q \in L_2$ and for all $p \in L_1$, (in such cases, $q \notin L_1$ for $p \leftrightarrow q$ holds true if $p=q$).

We compared pairwise entities and, for each comparison, the similarity degree d and set L of common (or, equivalent) ontological concepts are returned.

The comparison results are shown in Table 2 (*see*, Appendix for an example of comparison).

We defined inference rules to create a reasoner able to use the similarity values to automatically build the global conceptual schema [25]. The reasoner infers that the concept of person affiliated to the University in CWA corresponds to the concept of author in SAPERI. So, it is possible to build an integrated source relational schema where the publications are authored by persons affiliated to the University Structures according to their specific institutional role.

The global conceptual schema is then transformed into a relational schema by applying well-known rules [26]. The final schema is shown in Figure 8.

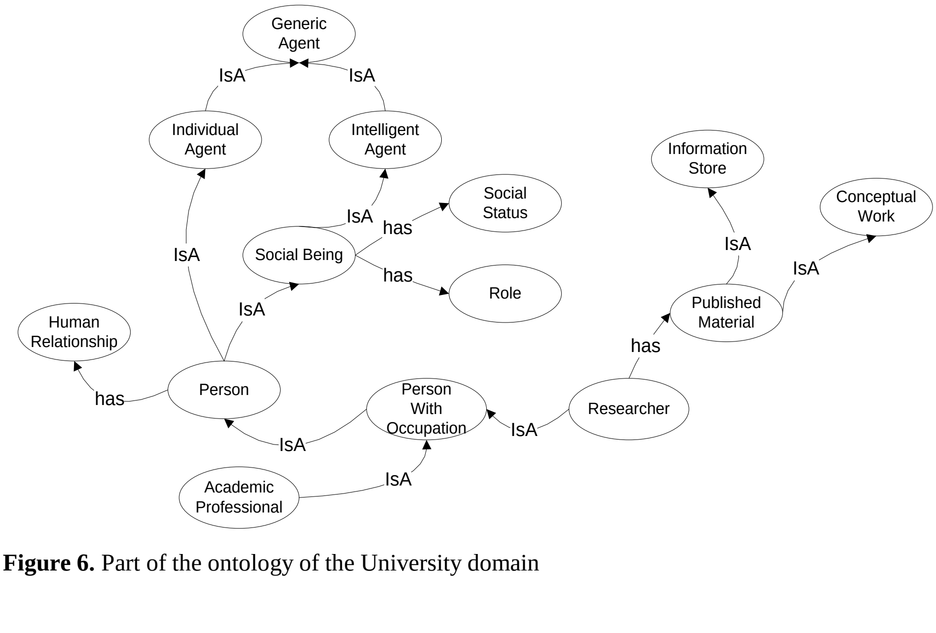


Figure 6. Part of the ontology of the University domain

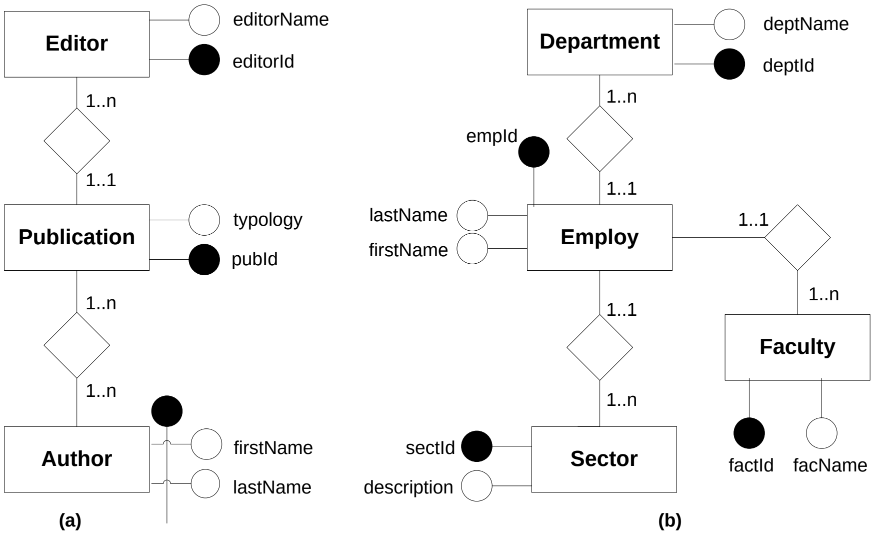


Figure 7. Source databases: (a) SAPERI; (b) CWA

Table 1. Entity definition

Database	Entity	Description	Entity definition
SAPERI	Editor	Company that published the paper	editor(X) ⇐ individualAgent(X) ∧ editorOfPublication(X,Y) ∧ publishedMaterial(Y)
SAPERI	Publication	Scientific paper published by the editor	publishedMaterial(X) ⇐ informationStore(X) ∧ conceptualWork(X)
SAPERI	Author	Person who works in the University and who authored the scientific paper	author(X) ⇐ personWithOccupation(X) ∧ academicProfessional(X) ∧ has(X,Y) ∧ publishedMaterial(Y)
CWA	Employ	Person who works in the University and is involved in research activities	employ(X) ⇐ personWithOccupation(X) ∧ academicProfessional(X) ∧ researcher(X) ∧ has(X,Y) ∧ publishedMaterial(Y)
CWA	Department	Research structure of the University	department(X) ⇐ researchOrganization(X) ∧ geographicalAgent(X)
CWA	Faculty	Teaching structure of the University	faculty(X) ⇐ educationalOrganization(X) ∧ academicOrganization(X)
CWA	Sector	Scientific area of a researcher	sector(X) ⇐ scientificFieldOfStudy(X)

Table 2. Similarity results

		CWA			
		Employ	Department	Faculty	Sector
SAPERI	Editor	0.34	0.22	0.22	0.26
		∅	∅	∅	∅
	Publication	0.19	0.25	0.25	0.29
		∅	∅	∅	∅
	Author	0.88	0.2	0.2	0.25
		{personWithOccupation(X), academicProfessional(X)}	∅	∅	∅

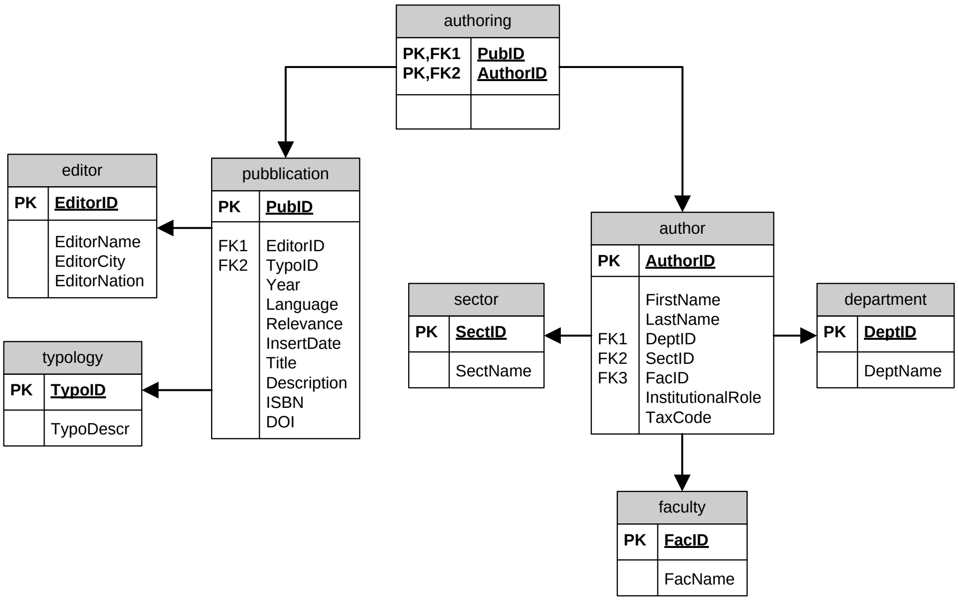


Figure 8. Part of the integrated source schema

4.3. Conceptual design

The conceptual design relies on the graph-based multidimensional model for representing the integrated global schema. In detail, this phase consists in a reengineering of the data source by performing traditional operations on graphs (such as prune, graft, and deleting nodes). However, the modelling process does not rely on the designer’s experience, but it is executed in automatic and supervised way, according to the requirements emerged from business goals [22].

In this case study, the constraint derived from the requirements emerged in Figure 5 states that we must have a publication cube, having no measure and a context of analysis composed of four dimensions: (i) author, having sector and department as further hierarchical levels; (ii) typology; (iii) editor; and (iv) year. So, we create a graph starting from the publication relation in Figure 8 and by navigating in the schema through foreign keys. The so-called attribute tree is shown in Figure 9a.

At this point, it is possible to remodel that attribute tree using an algorithm that applies the given constraint. To this end, the graph to be created must support the following guidelines in order to represent a multidimensional schema. The root node is the cube and the children of the root represent the measures of the fact table. The non-leaf nodes represent dimensional attributes, *i.e.*, entities that represent levels of aggregation. The number of dimensional attributes linked to the root establishes the dimensionality of the data cube. The dimensional attributes linked each other by an edge form a hierarchy. The leaf nodes represent the descriptive attributes of a dimensional attribute. The constraint imposes to introduce a year dimension as a root child. Then, a change parent operation is performed to make author as a further dimension, which has the new hierarchical level represented by institutionalRole. It is worth noting that a hierarchical level must have its descriptive attribute as a leaf node.

To summarize, the multidimensional schema of the Research Data Mart contains only one cube and is obtained from that integrated schema, by focusing the attention on the table containing the publications. The schema presents a four-dimension cube with no measures. Indeed, there are no numeric attributes related to a scientific publication. The first dimension is editor, the second dimension is typology, and the third one is time, whereas the minimum aggregation pattern for the time dimension is year. The last dimension is author, which is structured in four hierarchies: author → department, author → sector, author → faculty, and author → institutionalRole. Notice that the symbol → stands for a one-to-many relationship. Therefore, roll-up and drill-down operations are allowed. The final attribute tree is shown in Figure 9b.

Conceptual schema validation

The workload coming from requirements is now used in order to perform the validation process. If all the queries of the workload can be effectively executed using the schema, then such a schema is assumed to be validated and the designer can safely translate it into the corresponding logical schema. Otherwise, the conceptual design process must be manually revised.

We define the following issues related to the validation of a conceptual schema in reference to the queries included into the preliminary workload:

- a query involves a cube that has not been defined as such;
- a query requires a measure that is not an attribute of the given cube;
- a query presents an aggregation pattern on levels that are unreachable from the given cube;

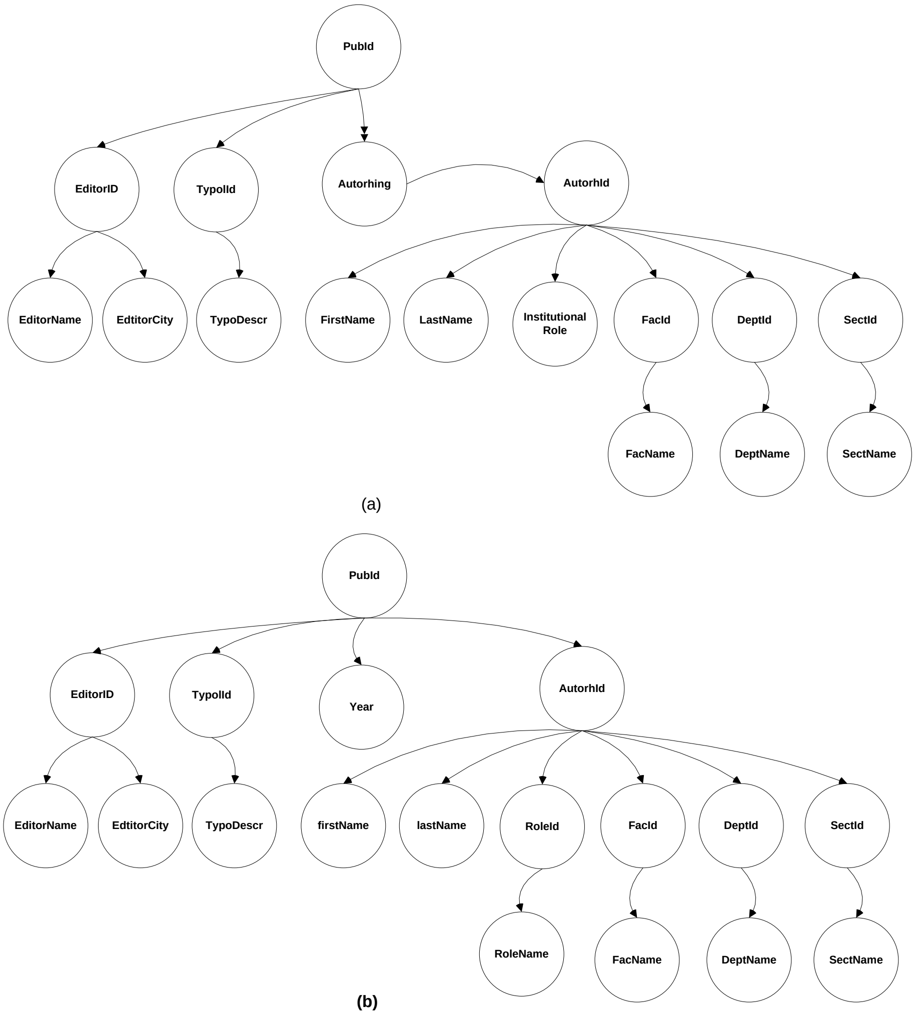


Figure 9. (a) attribute tree; (b) remodelled attribute tree

- a query requires an aggregation on a field that has not been defined as a dimensional attribute; and
- a query requires a selection on a field that has not been defined as a descriptive attribute.

A query is assumed to be validated if there exists at least an attribute tree such that the following conditions hold: (a) the fact is the root of the tree; (b) the measures are the children nodes of the root; (c) for each level in the aggregation pattern, there exists a path from the root to a node X, where X is a non-leaf node representing the level; and (d) for each attribute in the selection clause, there exists a path from the root to a node Y, where Y is a leaf node representing that attribute.

If all queries are validated, then each attribute tree can be considered as a cube. Then, we transform the cube so that the root is the fact table, non-leaf nodes are dimension tables, and leaf nodes are descriptive attributes belonging to a dimensional level.

In this case study, the root corresponds to the *publication* fact. Moreover, a non-leaf node exists for each aggregation level as defined in the requirements, whereas each level has got its own descriptive attributes. So, the conceptual schema satisfies the workload.

4.4. Logical design

In the logical design, the conceptual schema, formed of a set of one or more independent graphs, is transformed into a logical one based on the relational model, where the root node of the graph is a fact table and root children are dimension tables.

Figure 10 depicts the logical schema obtained from the attribute tree shown in Figure 9b. To complete the explanation of the design process, we also include the logical schema of the Didactics data mart (*see*, Figure 11).

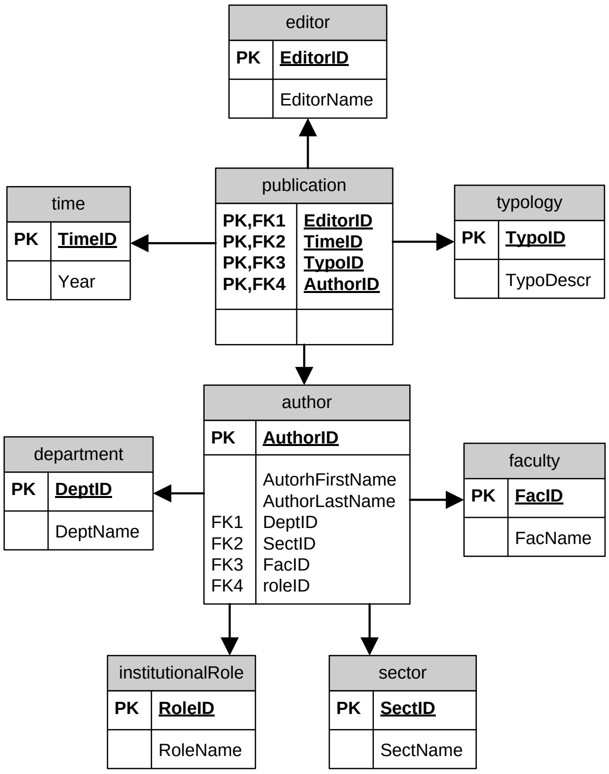


Figure 10. Research data mart

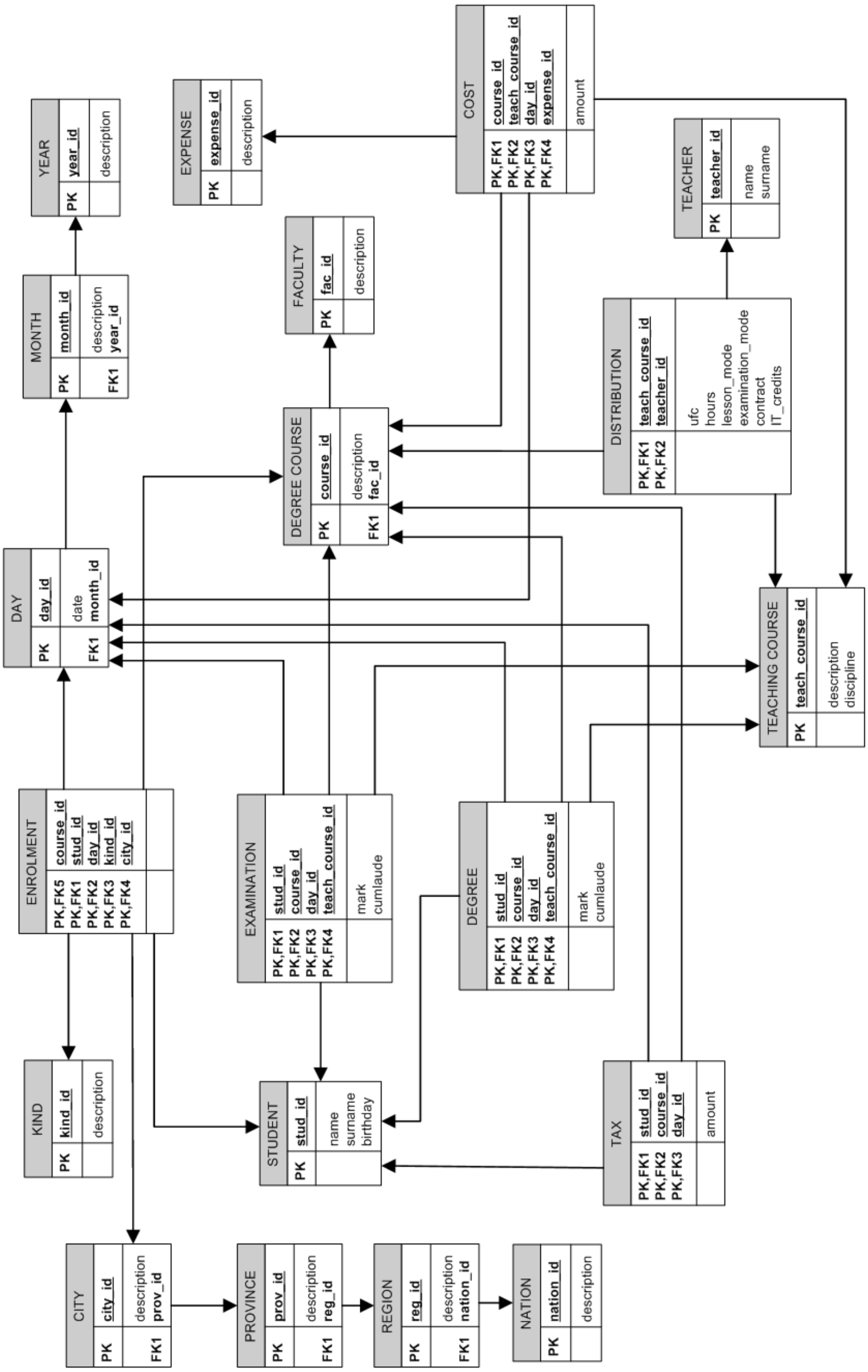


Figure 11. Didactics data mart

The Didactics data mart uses two source databases to load and refresh data: ESSE3– the main database–, and NOGE. It contains six fact tables: *enrolment, tax, examination, degree, distribution,* and *cost.* All these fact tables have three common dimensions: *student, degree course,* and *time*. These are basic dimensions for they represent the minimal information to express «who, where and when» aggregation levels.

The *enrolment* fact table has five dimensions. Here, the additional dimensions are: *residence*, that allows demographic or geographic aggregation, and *kind of enrolment*, that allows administrative aggregation. This cube has no measures and its function is to store the enrolment to a course of study by the student.

The *tax* fact table has only the *student*, *time* and *degree course* dimensions and it has *amount* as measure. Its function is to analyze the payment of the taxes by the student.

The *examination* fact table has four dimensions. Here, the additional dimension is represented by *teaching course*, that allows didactic aggregation. It has two fields as measures: the first field is the *mark*, that represents the fundamental measure to know students' performance; the second is the *cum laude* field, that is a simple Boolean field.

Also the *degree* fact table has four dimension tables, and it has the same additional dimension owned by the *examination* cube and measure and fields *mark* and *cum laude*.

The *distribution* fact table contains the list of teaching courses for each degree course of study and teacher. The measures include the number of teaching hours, the number of university formative credits (UFC), the kind of lesson, and the kind of examination.

The *cost* fact table is relative to the annual costs supported by the University for the management of each teaching course, and totally for each degree course per academic year. It also contains the teacher's costs for those teachers not enrolled in the University teacher's staff.

To obtain aggregate results at different levels of granularity, some dimensions are organized in dimensional hierarchy. In particular, the *degree course* dimension presents the *course* → *faculty* hierarchy, for allowing aggregate measures (*e.g.*, count of graduate students) at the degree study or faculty levels. The *residence* is a four-level dimensional hierarchy, for it presents the hierarchy *city* → *province* → *region* → *nation* for analyzing data according to different geographic contexts. Finally, the *time* dimension presents the three-level hierarchy *day* → *month* → *year* for summing data respectively by day, month, or year. All other dimensions of the Didactics data mart are one-level hierarchy.

4.5. Feeding process

ETL tools are systems to load data from source databases into target tables of the data warehouse. This feeding process requires a deep knowledge of the schema of the source databases, in order to properly map fields of source tables to those of target tables, and to store data. This process, which essentially addresses data integration, includes an important sub-process, whose aim is to perform data cleaning. In fact, this activity must be able to ensure a high-level quality of data, as these data will be used to provide information and knowledge for decision making.

To feed the academic data warehouse, we considered NOGE, ESSE3, CWA and SAPERI databases. During the design and the implementation of the ETL procedures, several problems arose due to inconsistencies among the source databases.

The source NOGE database contains dirty data due to input errors and the lack of controls by the software used for storing data. These errors frequently are null values or typos. However, severe errors are consistency errors or the presence of duplicated records. Therefore, in the ETL process, they occur two kinds of problems. The first one arises in populating dimensional tables. In fact, typical errors coming out when loading data into a dimension table are the violation of the primary key constraint. The second kind of problems regards the data mapping to fact tables. In this case, foreign key constraint violations occur frequently.

On the other hand, ESSE3 database does not present problems on referential integrity but contain problems due to data entry errors. The most important problem we faced relates to the authorship of publications. In fact, in the SAPERI database no constraints are defined on the author's name. As a consequence, users that insert data are allowed to associate any author's name to a paper, even if that name does not match with any person affiliated to the University. So, typos and misspelling are very frequent. Moreover, in each dataset, we also found incomplete names (without the forename), shortened names, and reversed names because of the lack of standardization in the name representation. Of course, these errors created homonyms problems and difficulties for the identification of the right author(s) of the paper. Some of these inconsistencies have been solved by comparing names with those in CWA and by using a data mining algorithm for string matching. Only a small part of the publications (the 5.8%) has not been associated to any person and it is gone lost. Furthermore, the 26.38% of the publications has been partially associated to legitimate authors, that is not all the authors have been correctly identified.

5. Data analyses

The analytical layer is represented by phpMyOLAP [27], an open source web application written in PHP and using MySQL as relational database management system, since this actually represents a valid solution for data warehousing environments [28]. Among the several storage engines provided by this system, we chose MyISAM, since this is a high performance engine. In fact, it is not transaction-oriented and it does not implement foreign key constraints. Indeed, in data warehousing systems, data consistency is more important than referential integrity [29].

PhpMyOLAP adopts the Mondrian XML schema format [30] to store the data warehouse metadata that can be browsed through a tree-based visualization. So, on the basis of the Query-By-Example approach, users can create reports without using the MDX language [31].

This application uses a native OLAP engine which supports traditional operators such as roll-up, drill-down, pivoting and generates SQL statements to be executed on MySQL. This engine does not rely on Java-based OLAP engine acting as a middleware and requiring further web servers as Tomcat. This makes phpMyOLAP independent and portable for Apache-MySQL-PHP systems.

5.1. Decision makers

The web application produces reports for university decision makers, who are internal or national agencies:

- *Academic supervisory staff*. There are two principal Academic supervisory staffs: the Academic Senate and the Administration Council. The Academic Senate is the governing body in matter of programming the development of the Athenæum and the coordination of Didactics and Research. It approves the criteria for the distribution of the financings among the Research Structures. Moreover, it determines the evaluation criteria of the didactic activities and estimates the effectiveness by analyzing the reports produced by the Evaluation Team. This is a partially elective independent team, named by the University Rector, to periodically verify the operating efficiency of all structures (the didactic structures, the research ones, and those for the technical-administrative management).
- *The Administration Council*. This team deliberates and supervises the administrative, financial, and economic-patrimonial management of the Athenæum. In particular, the Council deliberates about the performance of the criteria for the distribution of the financial resources among institutions and the technical and administrative staffs of the University.
- *Organizational structures*. Faculties are the fundamental structures that organize and coordinate the Didactic activities. In University, the management of the Research activities is entrusted to the Departments. The Departments are the organizational structures that collect teachers and researchers coming from several Faculties, but joined by common scientific interests and research methodologies. Departments collaborate with Faculties for the realization of the Didactic activities.
- *Administrative structures*. These structures are the student secretariats and the data elaboration centres, whose tasks are the production of data for the national "Alma Laurea" registry of the graduate students and the realization of documents, statements and other information reports to support the decisional processes.
- *National committee*. The national committee for the evaluation of the university system is an institutional team, whose tasks are: to establish the general criteria for the evaluation of the activities of the university; to predispose the annual report on the evaluation of the university system; to promote the experimentation, application, and spread of methodologies and evaluation tasks; to determine the nature of the information and data that the athenæum evaluation team must communicate; to predispose studies and documentation on the state of the university instruction, the compliance with the study right, and the accesses to the university courses of study.
- *Association of Rectors*. The Association of the Rectors of the Italian Universities, named *Crui*, was born in 1963 as a private association of the Rectors and, in short time, it has acquired a recognized institutional role and a concrete ability to influence the development of the university system through

an intense activity of study and experimentation. *Crui* centralizes its own evaluation activity in particular on the Didactics and Research areas, develops and proposes methodologies and evaluation criteria for athenæums, and degree courses, finalized to the improvement of the quality of the Italian university system.

5.2. Web application

The decision makers are interested in creating interactive reports by navigating through the schema on the basis of the multidimensional model. So, the starting point is a tree-based representation of each data mart. In Figure 12, we show the multidimensional elements of the Publication cube of the Research data mart. The example is devoted to compute the number of publications per Departments. The result of the analysis is shown in Figure 13. Decision makers are now allowed to execute traditional operations, such as order by the number of publications, and typical OLAP operations, such as roll-up/drill-down (*see*, Figure 14), slice-and-dice, drill-across, and pivoting.

Moreover, the application provides further features to public reports by sharing a link on the most popular social network [32] or by sending the link via email and to export results according to the *csv* or *pdf* formats.

The application can also produce statistics about the Didactics. The report in Table 3 considers historical data about the students' enrolments of years 2000 to 2005 per academic year and region. Here, the pivoting operator is applied to show the enrolments in reference to the regions in order to know which the most important affluence centres are for our University. We observe that, because residence dimension is a four-level dimensional hierarchy, the same analysis with a roll-up operation can produce a coarser grain result summing data for nation or, with the opposite drill-down operation, provide a finer-grained view considering the number of students at the province or city levels.

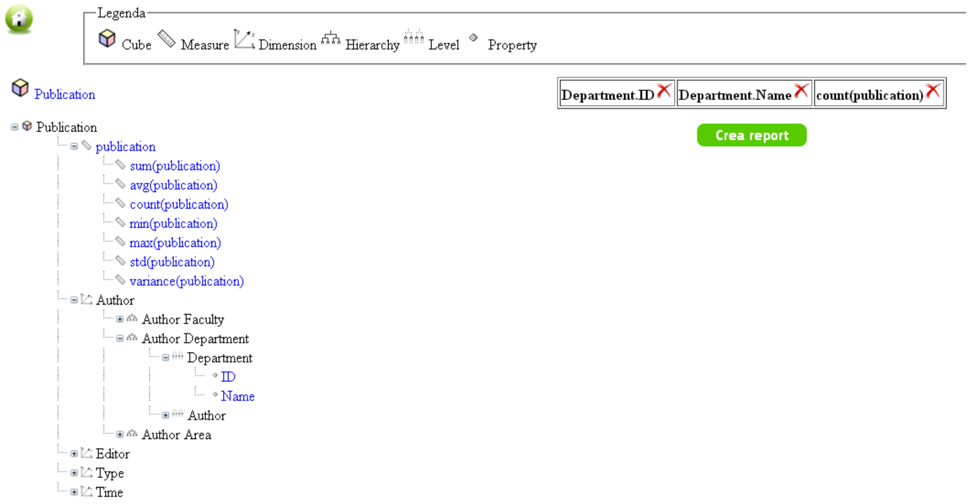
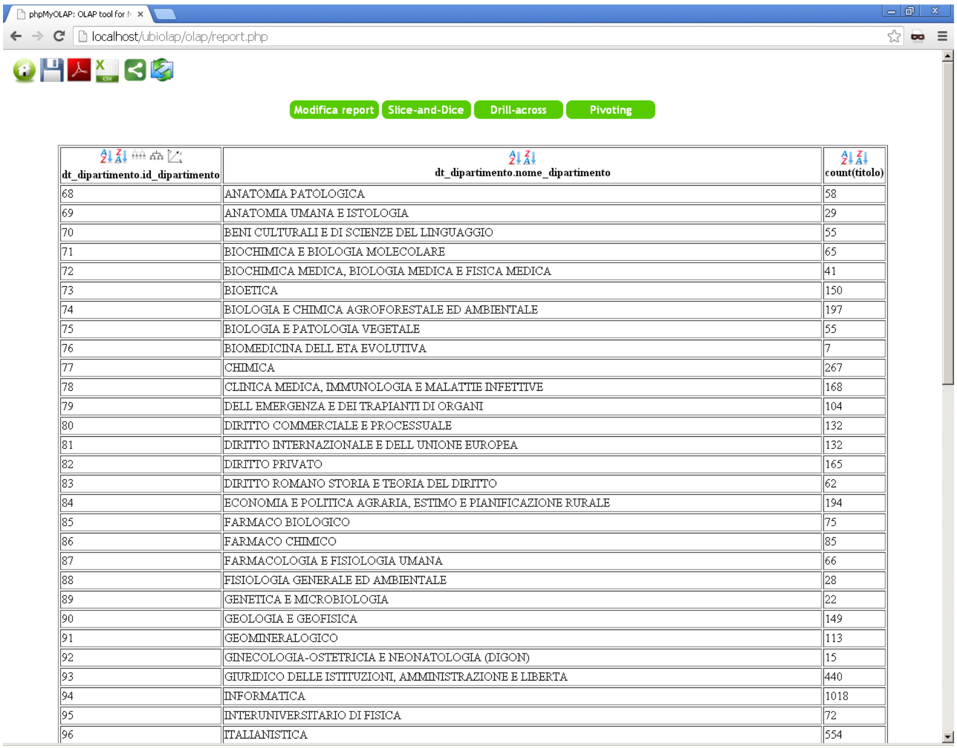
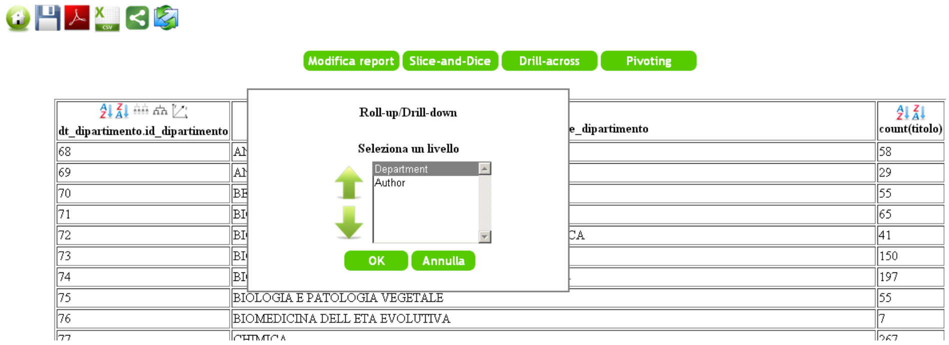


Figure 12. Research data mart



dt_dipartimento.id_dipartimento	dt_dipartimento.nome_dipartimento	count(titolo)
68	ANATOMIA PATOLOGICA	58
69	ANATOMIA UMANA E ISTOLOGIA	29
70	BENI CULTURALI E DI SCIENZE DEL LINGUAGGIO	55
71	BIOCHIMICA E BIOLOGIA MOLECOLARE	65
72	BIOCHIMICA MEDICA, BIOLOGIA MEDICA E FISICA MEDICA	41
73	BIOETICA	150
74	BIOLOGIA E CHIMICA AGROFORESTALE ED AMBIENTALE	197
75	BIOLOGIA E PATOLOGIA VEGETALE	55
76	BIOMEDICINA DELL ETA EVOLUTIVA	7
77	CHIMICA	267
78	CLINICA MEDICA, IMMUNOLOGIA E MALATTIE INFETTIVE	168
79	DELL EMERGENZA E DEI TRAPIANTI DI ORGANI	104
80	DIRITTO COMMERCIALE E PROCESSUALE	132
81	DIRITTO INTERNAZIONALE E DELL UNIONE EUROPEA	132
82	DIRITTO PRIVATO	165
83	DIRITTO ROMANO STORIA E TEORIA DEL DIRITTO	62
84	ECONOMIA E POLITICA AGRARIA, ESTIMO E PIANIFICAZIONE RURALE	194
85	FARMACO BIOLOGICO	75
86	FARMACO CHIMICO	85
87	FARMACOLOGIA E FISIOLOGIA UMANA	66
88	FISIOLOGIA GENERALE ED AMBIENTALE	28
89	GENETICA E MICROBIOLOGIA	22
90	GEOLOGIA E GEOFISICA	149
91	GEOMINERALOGICO	113
92	GINECOLOGIA-OSTETRICIA E NEONATOLOGIA (DIGON)	15
93	GIURIDICO DELLE ISTITUZIONI, AMMINISTRAZIONE E LIBERTA	440
94	INFORMATICA	1018
95	INTERUNIVERSITARIO DI FISICA	72
96	ITALIANISTICA	554

Figure 13. Report on the number of publications per Department



dt_dipartimento.id_dipartimento	e_dipartimento	count(titolo)
68	AI	58
69	AI	29
70	BE	55
71	BI	65
72	BI CA	41
73	BI	150
74	BI	197
75	BIOLOGIA E PATOLOGIA VEGETALE	55
76	BIOMEDICINA DELL ETA EVOLUTIVA	7
77	CHIMICA	267

Figure 14. Roll-up/Drill-down on the Department hierarchy

The report in Table 4 groups the same data by academic year and university Faculty. In this case, since degree course is a two-level dimensional hierarchy, the same analyses with a drill-down operation provide a detailed map showing the counts of students grouped by degree course.

Table 3. Count of enrolled students grouped by Academic year and Region

REGION	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**
Puglia	64396	65037	70529	68542	60674	60499
Basilicata	3264	2945	3088	2824	2460	2691
Calabria	813	739	778	778	649	719
Greece	306	250	237	219	155	102
Lombardia	135	106	104			
Lazio	129		119	103		
Campania	123	129	208	260	175	332
Molise			221	136	106	
Sicilia			114	127		163

Table 4. Count of enrolled students grouped by Academic year and Faculty

Faculty	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**
Law	15222	13092	12634	11408	8619	9026
Economics	9935	9797	10235	9254	7496	7904
Educational Sciences	8920	9814	11187	12346	13064	11963
Mathematics, Physics and Natural Sciences	6806	8525	9821	8907	6795	6622
Medicine and Surgery	5659	5427	6345	7075	7464	8080
Arts and Philosophy	5409	4792	5675	5288	4506	4920
Political Sciences	4732	4388	4598	4268	3092	3417
Pharmacy	3786	4242	3550	3700	3984	4430
Foreign Languages and Literatures	3318	3532	4304	4364	3841	3833
Law (Taranto city)	1816	2148	2806	2862	2278	2164
Veterinary Medicine	1431	1332	1592	1744	1835	1799
Agricultural Sciences	1296	1079	1530	1347	938	821
Economics (Taranto city)	901	801	745	689	584	693

6. Conclusion

In the paper, we showed the system architecture of a Business Intelligence system for academic organizations. The core of the system is a data warehouse which allows university decision makers to analyze crucial aspects related to the evaluation of the quality of both Didactics and Research. The most difficult part of the realization of the data warehouse is represented by the design process that must consider simultaneously both different data sources and user requirements. To this end, we adopted a hybrid methodology that is largely automatic and able to integrate different data sources on the basis of an ontological approach. We encountered also several problems in the feeding process, due to data entry errors and lack of constraints in legacy systems. Business applications are developed and deployed using a web-based OLAP layer, which has been released as an open source project that offers also the possibility to share reports using social networks. So, users can perform a collaborative analysis of data, by posting comments and opening discussion on the results. Future work will devoted to extend the methodology in order to include also agile techniques for reacting in a fast way to frequent changes in academic regulations that imply new types of analysis.

References

1. Akintola K.G., Adetunmbi, A.O. and Adeola O.S.: Building Data Warehousing and Data Mining from Course Management Systems: A Case Study of FUTA Course Management Information Systems. International Journal of Database Theory and Application, 4(3), 13-24. (2011)
2. dell'Aquila C., Di Tria F., Lefons E. and Tangorra F.: An Academic Data Warehouse. In Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications: WSEAS Press, 229-235. (2007)
3. Tanuska P., Vlkovic O., Vorstermans A. and Verschelde W.: The proposal of ontology as a part of University data warehouse. In Proceedings of the 2nd International Conference on Education Technology and Computer, volume 3, 21-24. (2010)
4. Muntean M., Bologa A., Bologa R. and Florea A.: Business Intelligence Systems in Support of University Strategy. In Recent Researches in Educational Technologies, 118-123. (2011)
5. Lin M.C.: University data warehouse design issues: case study. In Proceedings of the 2001 American Society for Engineering Education Annual Conference & Exposition, 1-9. (2001)
6. Donhardt G.L. and Keel D.M.: The Analytical Data Warehouse: Empowering Institutional Decision Makers. Educause Quarterly, 24(4), 56-58. (2001)
7. Fernandes C. and Whalen M.: Data Warehousing from the Web. In Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition, 1-11. (2004)
8. Ćamilović D., Bečejski-Vujaklija D. and Gospić N.: A Call Detail Records Data Mart: Data Modelling and OLAP Analysis. Computer Science and Information Systems, 12, 87-110. (2009)
9. Di Tria F., Lefons E. and Tangorra F.: Research Data Mart in an Academic System. In 2012 Spring Congress on Engineering and Technology, IEEE, 18-22. (2012)
10. Di Tria F., Lefons E. and Tangorra F.: Hybrid methodology for data warehouse conceptual design by UML schemas. Information & Software Technology, 54(4), 360-379. (2012)
11. Romero O. and Abelló A.: A Survey of Multidimensional Modeling Methodologies. International Journal of Data Warehousing and Mining, 5, 1-23. (2009)
12. Phipps C. and Davis K.C.: Automating Data Warehouse Conceptual Schema Design and Evaluation. In Proceedings of the 4th International Workshop Design and Management of Data Warehouses, 23-32. (2002)
13. Triola M. and Pusic M.: The Education Data Warehouse: A Transformative Tool For Health Education Research. Journal of Graduate Medical Education, 4(1), 113-115. (2012)
14. MedBiquitous Consortium. http://www.medbiq.org (accessed November 15, 2011).
15. Dewitt J.G. and Hampton P.M.: Development of a data warehouse at an academic health system: knowing a place for the first time. Acad Med, 80(11), 1019-25. (2005)

16. Einbinder J.S., Scully K.W., Pates R.D., Schubart J.R. and Reynolds R.E.: Case study: a data warehouse for an academic medical center. J Healthc Inf Manag, 5(2), 165-75. (2001)
17. Deniz D. and Ersan I.: An Academic Decision-Support System Based on Academic Performance Evaluation for Student and Program Assessment. Int. J. Engng Ed., 18(2), 236-244. (2002)
18. dell'Aquila C., Di Tria F., Lefons E. and Tangorra F.: Dimensional fact model extension via predicate calculus. In Proceedings of the 24th International Symposium on Computer and Information Sciences, IEEE, 211-217. (2009)
19. Di Tria F., Lefons E. and Tangorra F.: GrHyMM: A Graph-Oriented Hybrid Multidimensional Model, In Advances in Conceptual Modeling. Recent Developments and New Directions, Lecture Notes in Computer Science, vol. 6999: Springer, 86-97. (2011)
20. Mazón J.N., Trujillo J., Serrano M. and Piattini M.: Designing Data Warehouses: from Business Requirement Analysis to Multidimensional Modeling. In Requirements Engineering for Business Need and IT Alignment; University of New South Wales Press, 44-53. (2005)
21. Di Tria F., Lefons E. and Tangorra F.: Ontological Approach to Data Warehouse Source Integration. In Proceedings of the 28th International Symposium on Computer and Information Sciences, Springer, 251-259. (2013)
22. dell'Aquila C., Di Tria F., Lefons E. and Tangorra F.: Logic Programming for Data Warehouse Conceptual Schema Validation. In Data Warehousing and Knowledge Discovery, the 12th International Conference, Springer, 1-12. (2010)
23. Reed S.L. and Lenat D.B.: Mapping Ontologies in Cyc. In AAAI 2002 Conference Workshop on Ontologies for the Semantic Web, Edmonton, Canada. (2002)
24. Ferilli S., Basile T.M.A., Biba M., Di Mauro N. and Esposito F.: A General Similarity Framework for Horn Clause Logic. Fundam. Inform, 90(1-2), 43-66. (2009)
25. Di Tria F., Lefons E. and Tangorra F.: Big Data Warehouse Automatic Design Methodology. In W. Hu, & N. Kaabouch (Eds.) Big Data Management, Technologies, and Applications, Hershey, PA: Information Science Reference, 115-149. (2014**)**
26. Elmasri R. and Navathe S.B.: Fundamentals of Database Systems, 6th Edition, Addison-Wesley. (2010)
27. PhpMyOLAP. http://phpmyolap.sourceforge.net
28. Feinberg D. and Beyer M.A.: Magic Quadrant for Data Warehouse DBMS Servers: Gartner, (accessed February 2014 from Sybase Web site: http:// www.sybase.com). (2011)
29. Golfarelli M. and Rizzi S.: Data Warehouse Design: Modern Principles and Methodologies, McGraw-Hill/Osborne Media. (2009)
30. Mondrian documentation. http://mondrian.pentaho.com/documentation/schema.php
31. Spofford G.: MDX Solutions, John Wiley & Sons Inc. (2001)
32. Zhao Du, Xiaolong Fu, Can Zhao, Ting Liu and Qifeng Liu.: University Campus Social Network System for Knowledge Sharing. Computer Science and Information Systems, 9(4), 1721-1737. (2012)

Appendix

Here we show the comparison process between the entities *author* of *SAPERI* and *employ* of *CWA*. These entities were previously defined as (*cf.*, Table 1)

author(X) ⇐ personWithOccupation(X) ∧ academicProfessional(X) ∧
has(X, Y) ∧ publishedMaterial(Y),

employ(X) ⇐ personWithOccupation(X) ∧ academicProfessional(X) ∧ researcher(X) ∧
has(X, Y) ∧ publishedMaterial(Y).

First, we bind the variable *X* of *author* to the variable *X* of *employ*. Then, we create lists L_1 and L_2 using the predicates present in the respective logical definition. We informally have:

L_1 = {personWithOccupation(X), academicProfessional(X), has(X, Y), publishedMaterial(Y)},

L_2 = {personWithOccupation(X), academicProfessional(X), researcher(X), has(X, Y), publishedMaterial(Y)}.

To do the comparison between predicates, say, *p* and *q*, we introduce the mapping operator ↔ defined so:

$$p \leftrightarrow q = \begin{cases} 1 & \text{if } p = q \text{ or they have a common generalization in the ontology} \\ & \textit{(successful mapping)} \\ 0 & \text{if } p \text{ and } q \text{ have no ontological relationship} \\ & \textit{(failure)} \end{cases}$$

In case of successful mapping on *X*, the common or superior concept is added to the generalization list *L if not present* and counter *l* increases by 1. In case of unsuccessful mapping, they opportunely increase *n* and/or *m*. Notice the mapping operator is commutative (*i.e.*, $p \leftrightarrow q = q \leftrightarrow p$).

Then, considering one bound variable at a time, we compare first each unary predicate in L_1 with every unary predicate in L_2, after that the binary predicates, and so forth.

Mapping 1. personWithOccupation(X) ↔ personWithOccupation(X) = 1 for predicates can be successfully mapped. Therefore, one common predicate has been found, and *l* increases by 1. We have the partial results
L = {personWithOccupation(X)}, $l = 1$, $n = 0$, and $m = 0$. □

Mapping 2. personWithOccupation(X) ↔ academicProfessional(X) = 1. In fact, personWithOccupation is a generalization of academicProfessional. The superior predicate is already present in *L*, and *l* increases by 1. So,
L = {personWithOccupation(X)}, $l = 2$, $n = m = 0$. □

Mapping 3. Similarly, personWithOccupation(X) ↔ researcher(X) = 1. Thus,
L = {personWithOccupation(X)}, $l = 3$, $n = 0$, and $m = 0$. □

Mapping 4. academicProfessional(X) ↔ personWithOccupation(X) is the symmetric of mapping 1. So, we skip it. □

Mapping 5. academicProfessional(X) ↔ academicProfessional(X) = 1. Thus,
L = {personWithOccupation(X), academicProfessional(X)}, $l = 4$, and $n = m = 0$. □

Mapping 6. academicProfessional(X) ↔ researcher(X) = 1 for they both are specializations of personWithOccupation. Hence,
L = {personWithOccupation(X), academicProfessional(X)}, $l = 5$, and $n = m = 0$. □

We now consider the bound variable *Y*.

Mapping 7. Although publishedMaterial(Y) ↔ publishedMaterial(Y) = 1, no predicate is added to *L* because we are looking for the common concept(s) between *author* and *employ* , Then,

L = {personWithOccupation(X), academicProfessional(X)}, $l = 6$, and $n = m = 0$. □

Since unary predicates bound to all variables are terminated. we proceed with comparing each binary predicates in L_1 with those in L_2 relative to *X*.

The only binary predicate in both lists L_1 and L_2 is has(X,Y). In order to explicitly show this binary relationship and the involved entities, we consider the bound variables *X* and *Y* and make the substitutions:

(i) has(author, publishedMaterial)
that means that an uthor is an academic person having publications, and

(ii) has(employ, publishedMaterial)
that means that an employ is a researcher having publications.

So,

Mapping 8. has(author, publishedMaterial) ↔ has(employ, publishedMaterial) = 1 for entities *author* and *employ* have common ontological concepts (*i.e.*, $L \neq \emptyset$). Finally,

L = {personWithOccupation(X), academicProfessional(X)}, $l = 7$, $n = 0$, and $m = 0$. □

Therefore, $d = 0.875$ is the similarity degree and L = {personWithOccupation(X), academicProfessional(X)} is the list containing the common concept(s). Since 0.7 is the similarity threshold and $L \neq \emptyset$, the two entities *author* and *employ* refer to the same ontological concept for $0.875 > 0.70$ (*cf.*, Table 2).

Francesco Di Tria received his Laurea degree in Informatics from University of Bari “Aldo Moro” in 2007. In 2011, he received the PhD in Informatics from the same University. He is currently a research fellow in the Computer Science Department, where he teaches courses of “Databases”, “Information Systems”, and “Computer Network”, for the computer science curriculum. His research interests are in data warehousing and business intelligence systems. In detail, his main activity regards the automation of processes for data warehouse modelling and the definition of metrics for data warehouse quality measurement.

Ezio Lefons is Associate Professor of Informatics and teaches the “Data Bases” and “Advanced Data Bases” courses for the computer science curricula. His scientific research is relative to the field of data bases and, in particular, to database query systems based on finite multi-valued logic. These studies have lead to obtain International Patents. In the last years, the research has been oriented to the analytical database models, decision support systems, and approximate query systems in data warehouse environments.

Filippo Tangorra received the “Laurea” degree in Computer Science from University of Bari, Italy, in 1975. He is currently an associate professor at the Computer Science Department of the University of Bari, Italy, where he teaches courses of “computer architecture”, “database systems” and “information systems” for the computer science curriculum. His research was initially in human relational interaction system modelling. Successively, his research activity has been oriented to decision support systems, and statistical databases database models, standards for data dictionary systems (IRDS), and query optimization. His research interests in computer architecture led also to the definition of a processor simulation environment for teaching purpose. Current research concerns synopses for massive data in approximate query processing context, methodologies for data warehouse design, and data warehouse quality metrics definition. On these topics, he has published more than 100 papers.

Received: March 25, 2014; Accepted: August 20, 2014