Merging graduate data from remote and non-homogeneous information systems

Alberto Leone[1], Angelo Guerriero[2], Loris Cancellieri[3], Piero Di Sario[4]

Consorzio Interuniversitario AlmaLaurea, Viale Masini 36, 40126 Bologna (BO), Italy
[1]alberto.leone@almalaurea.it, [2]angelo.guerriero@almalaurea.it, [3]loris.cancellieri@almalaurea.it
[4]piero.disario@progel.it

Keywords
Graduate database, ETL, data warehouse

1. EXECUTIVE SUMMARY

The AlmaLaurea Interuniversity Consortium, founded by the University of Bologna in 1994, currently gathers 51 Italian Universities (i.e. 70% of the graduate population) and mainly aims at creating and maintaining a database of all graduates from its member Universities. This database originates from the administrative data of all graduates, as supplied by the individual Universities to the central database in certain times of the year. Administrative data are matched with the information collected through an online questionnaire. These data form the complete database which is used for annual statistical surveys. Part of the data are used for creating a CV which is accessible online by the companies which are recruiting graduates for their staff.

1.1. Background

All Italian Universities process their students' data using management software systems. Even though some of these systems are based on standard software solutions (Kion, Cineca), in most cases software solutions are specially developed within each individual University. Each of these systems is based on a special database structure, on a specific set of collected and managed information and on dedicated encoding schemes for the structured variables. Furthermore, all data management systems are designed and developed with a view to simplifying data entry operations and the way in which students' career records are kept. The goals and design constraints of such systems do not always take into account the overall database consistency which is supposed to be a key prerequisite for the creation of a data warehouse. The setting up of a central database – which has been in place since 1994 – therefore required a solution to the above problems. Another major design constraint was the need to keep the data structure compatible over time without it being affected by the continuous changes imposed by the developing regulations (Bologna process) and by the routine update of the information systems used by each University.

1.2. The solution

The solution adopted by AlmaLaurea relies on two basic elements:

1. development of a data exchange format which has changed in time from fixed spacing to xml, ensuring compatibility with the previous formats.

2. implementation of a software application which handles data extraction, processing and loading (ETL)

Before being finally uploaded into the system, all data go through a delicate staging phase where the staff of AlmaLaurea and of the individual Universities closely cooperate to identify and correct any error and inconsistency entered in the dataset at the time of loading.

The aim of this paper is to provide a detailed description of the complex ETL system implemented by AlmaLaurea: from the data exchange format to the highly parameterized solution for data verification and editing, all data processing stages from data extraction to final publication in the online database will be discussed.

2. The AlmaLaurea Interuniversity Consortium

AlmaLaurea is an innovative service that puts graduate CVs and résumés online, thus serving as a meeting point between graduates, universities and companies. Founded in 1994 on the initiative of the Statistical Observatory run by the University of Bologna, AlmaLaurea has grown exponentially and today covers every year 67% of Italian graduates.

Managed by a consortium of Italian universities with the support of the Ministry of Education, University and Research, AlmaLaurea was set up with the aim of putting businesses and graduates in contact and establishing itself as a reference point within the university system for the subjects involved (scholars, operators, etc.) in university education, employment and the development of young people in general.

Every year AlmaLaurea publishes two analyses on Italian graduates:

Annual Graduate Profile Report.

The Report examines all the graduates of the year, considering their characteristics and performances in the light of a multitude of variables including age at graduation, continuity of studies and attendance, parents' education, social background, study abroad, apprenticeships or internships, foreign languages and IT skills, etc.

Annual Report on the Occupational Conditions of Graduates.

The Report provides in-depth information about the employment conditions of young graduates at one, three, and five years from completion of studies, the prospects of the labour market and the relationships between university studies and employment opportunities.

These two reports are available in both paper and digital version. See References.

3. Problem issues

In general terms, the problem that is about to be dealt with boils down to a data extraction, uploading and processing procedure (ETL) typical of data warehouse creation. Usually, the acquisition of data that are produced and processed for management purposes only presents with inherent quality limitations when these data need to be utilized for statistical purposes in a Business Intelligence system.

By way of example, consider an open-text entered field such as the description of a place. From a merely functional point of view, a place like, for instance, "SAN LAZZARO DI SAVENA", an Italian small town located near Bologna, might be indicated in different ways, without leading to functional errors. In other words, if someone were to dispatch correspondence to an address located in that place, one could indifferently use strings like:

```
SAN LAZZARO DI SAVENA
S. LAZZARO DI SAVENA
S. LAZZARO
SAN LAZZARO
```

They would all be valid and the message would be delivered in any case (in this specific example, the postcode should also be indicated).

Conversely, if one were to count how many students have graduated in a given year, or how many are enrolled on a specific course, broken down by place of residence, one would obtain a wrong count scattered across the diverse synonyms of the relevant variable.

More generally, the quality and integrity limitations of data defined on management information systems are on the whole different and, from a number of points of view, more flexible than those imposed upon by a data warehouse.

Hence data need cleaning and renormalizing before being entered.

This problem, which is recurrent in data warehouse feeding processes, is further compounded by another one which affects in particular the context where AlmaLaurea is called upon to operate, i.e.

the fact that databases where data are extracted from are very differentiated and inhomogeneous, such differences existing in the logical organization, structure and encoding schemes.

3.1. Complexity of the scenario

From the point of view of the information system in use, the scene of Italian universities is fairly variegated, even though, in the wake of a number of initiatives started by the Ministry, such as the registry of students and the database of training and education supply, a number of standardizations have been recently introduced.

A number of universities adopted the standard solution offered by Kion (a company incorporated and fully owned by the Cineca Interuniversity Consortium – www.cineca.it), thus significantly simplifying the problem. Nevertheless, the vast majority of universities resort to proprietary solutions with data structures, encoding and processes that are strongly customized and diversified.

This is further compounded by the growing complexity arising from the evolution of the regulatory framework which has experienced, for example, the passing of two significant pieces of legislation in as little as five years:

- In 1999, with Ministerial Decree 509/1999 (MIUR, 1999), the reform of the university system was set up, in compliance with the guidelines of the Bologna Process, envisaging the introduction of two degree levels (known as 3+2) and the reorganization of degree courses compelled to fall within predefined templates regarding the subject matters (degree course classes). University began to implement the new provisions in 2001.
- In 2004, by means of Ministerial Decree 270/2004 (MIUR, 2004), a change in the set-up of degree courses was introduced, with the redefinition of degree classes. A number of universities began to enforce the new regulatory framework only as late as in 2007, progressively applying the new regulations that are planned to be fully in force by 2010/2011.

The Italian university system provides the students with a remarkable degree of flexibility in terms of organization and attendance of degree courses. As a result, today there are a number of university students that attended courses organized under the pre-reform system (prior to 2001) who are still graduating. In general, the adoption of the new regulatory framework is gradual and inhomogeneous, even within single universities.

Practically speaking, in the next few years, each university will be expected to manage university careers under at least three different regulatory frameworks: pre-reform, post-reform 509 and post-reform 270.

Such complexity significantly affects the 51 universities members of the Consortium, when information on graduates is gathered in one single standardized central database used for statistical purposes.

4. Data processing in AlmaLaurea

4.1. The solution

The problem has been tackled by AlmaLaurea ever since the beginning of its activity, in 1994, when it faced the issue of the collection of data from the universities of the Emilia-Romagna Region.

The solution adopted at that time, though updated in terms of formats and technological solutions, is still in use in spite of the huge increase in volumes of data and universities involved.

The main ingredients of the solution are the definition of a standardized data exchange format and a semi-automated procedure for the verification of quality and correction of errors according to a set of predetermined rules.

Data control and verification are entirely performed via SQL queries such as

```
For each K-rule
SELECT *
FROM tab1, …, tabN
WHERE COND[k](field1_1,…,fieldM_1,…., fieldM_N)
       AND ID_RECORD NOT IN (SELECT ID_RECORD
                             FROM FORCED
                             WHERE FORCED_RULES = COND[K]
                             )
```

Each query responds with a Qk record, one per violation. The violation can be removed by correcting its data or by forcing its validity.

The solution is comprised of three main parts:

- The AlmaOne database. This is the throbbing heart of the entire system. It can be considered a meta-database which, besides memorizing administrative records and questionnaire data, contains all the domain and consistency rules that data have to comply with. A log-file is kept of all changes made by either the automatic scripts or the users. One of AlmaOne's fundamental parts is the data dictionary, that is a table containing all the information on the names of the fields, their type, format, codings, and inner tables where these are stored.
- The SWAL application. This is the acronym of SoftWare AlmaLaurea, and is the application that has witnessed the larger number of technological evolutions from 1994 to date. Started as a simple Microsoft Access database containing a set of queries for the preliminary control of data, it subsequently turned into a desktop application distributed to the universities and currently is a powerful web-based extranet system for the uploading, control and correction of data using Microsoft .NET, AJAX and Web-Services technologies.
- The Alma Flow application. This is used to guide users in the control and correction of administrative data and the matching of these two data sources so as to create the graduate's CV. One of the distinctive features of this application is its functioning by sets of data, known as worksets; in every moment all data related to a CV belong to one single workset. The use of worksets generates an horizontal partitioning of data thus permitting to have all data of different universities in the same database without any possibilities to get them muddled. The technology used is Microsoft ASP and is accessible only to some staff members.

Before moving on to a detailed analysis of the solutions adopted, an overview of the "data cycle" in the AlmaLaurea system is provided (see the scheme in Fig. 1).

AlmaLaurea processes data coming from two sources: universities send graduates' administrative data (personal data, academic qualification completed, university career, degree dissertation, etc) whereas each graduate, filling in the online questionnaire, provides information on his/her address(es), traineeships or study abroad experiences, if any, characteristics of the job sought and a number of self-assessments on foreign language and IT skills. Data originating from either source are uploaded on the AlmaOne database and undergo a first correction stage, subsequently they are matched so as to generate the graduate's CV. The data contained in the CV are in turn submitted to a set of consistency checks, upon completion of which data are published. During this phase, data are extracted from AlmaOne and uploaded on the dbCVTOT database which contains all the graduate CVs and is accessible via the web by means of the query interfaces available on the AlmaLaurea website. From that moment on, graduates can access their CVs online by means of the credentials supplied on registration and can modify the information contained, also specifying work experiences and postgraduate training.

Subsequent to publication, CVs are also uploaded on AlmaLaurea's data-warehouse where data are combined with all other available information (such as data on web traffic, mailing, accesses of graduates, purchases of CVs by companies, database searches, etc). These data are utilized exclusively for statistical purposes, under anonymity, and are disseminated in aggregate form only. This system is supported by an application which, on a regular basis, downloads from the online database the data of updated CVs and aligns the information present on the data-warehouse.

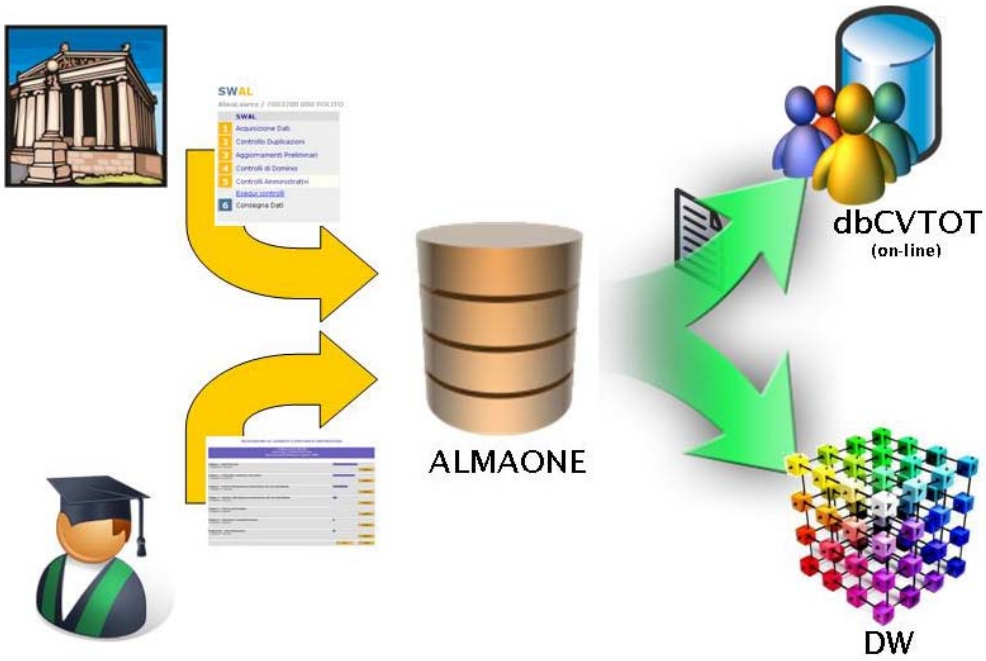

Fig. 1. DataFlow in the AlmaLaurea system

4.2. Acquisition phase

Extraction of data from university records

Administrative data on graduates are stored by different universities in different formats and the first step made was the definition of a data exchange standard that all universities had to comply with. The data exchange standard is known as AlmaLaurea administrative layout, which is comprised of a set of fields that universities are required to convey, their codings, the type and maximum size of data contained by such fields. While the layout is fixed and its constraints must be complied with, more freedom is granted to the user concerning the exchange file format. The system that has been created accepts file in XML format and, to guarantee compatibility with extraction processes still used by universities, also in formats with fixed spacing and variable spacing with separating character.

Data upload and grammar check

The resulting file is subsequently uploaded in AlmaLaurea's servers via the SWAL application upload page. After the upload, a number of checks on the file are performed, verifying that basic type constraints defined in the data layout are respected and that all compulsory fields are filled in for all the records. If the outcome is positive, the file is processed via an SQL Server DTS (Data Transformation Services) package and uploaded on a staging table.

Data duplication checks

Once the file has been accepted, duplication checks are performed in order that the presence of duplicated records may be detected both within the file and in the entire database. The check may be effected by name, date of birth, graduation and enrolment date. Double records already present

in the database and published are automatically erased, whereas, if duplicates are detected in the file, a screenful is displayed with a summary of record data and the user is asked to manually select the one to maintain.

At the end of this phase, data are transferred from one single staging table to the normalized tables of the AlmaOne database. Data are allocated to a specific workset known as *Officina* (workshop) and remain there until completion of the first checking and correction phase which is performed directly by the staff responsible in each university.

Domain belonging checks

Each attribute may be bound to take one predetermined values or values limited by specific ranges.

Before giving a detailed description of check phases, an overview of checks and operation modes will be provided. Checks consist of a set of queries that are operated on data to check for their correctness (domain checks) and their consistency (administrative checks). Domain checks can be considered as deferred referential integrity checks (Cammarata, 1989). In comparison to classical integrity checks of the relation model, they have some features that lack in many DBMS:

- Data are accepted by the system even though they don't complied with constraints
- In any moment you can know which records are invalid and you can correct them
- Only valid records can be admitted to the next stage of processing

Domain checks are comprised of a set of queries that are dynamically built based on the information contained in the data dictionary and that return to the user the records containing fields encoded with values that are not present in the corresponding decoding tables. The typical query is as follows

```
SELECT *
FROM Table
WHERE Field IS NOT NULL AND NOT EXISTS (
                                        SELECT *
                                        FROM Domain_Table
                                        WHERE Domain_Table.Code = Table.Field
                                        )
```

Where the Table, Field, Domain_Table, Code parameters are retrieved from the data dictionary record relevant to the field being checked.

Semantic check

Unlike the previous checks, administrative checks involve more than one field in one or more tables and are aimed a checking consistency in the different fields. This library of checks is continuously evolving and results from the experience accrued by AlmaLaurea in the data checking process. Administrative checks are stored in a specific AlmaOne table and can be inputted via either an SQL query or a dedicated AlmaFlow form.

Data that are necessary for the definition of administrative checks are the fields to be displayed on the data correction grid and the condition identifying potentially erroneous records (*where* clause). Starting from these data, the application engine retrieves from the dictionary the check-relevant tables and automatically generates the query also including the necessary join conditions. In the rule definition there is a possibility to specify the gravity level; it is possible to decide whether the records selected by the check should be classified as "erroneous" (high gravity) or, more simply, as "suspicious" (low gravity). Administrative checks are split into two subsets: those active only in SWAL and those active only in AlmaFlow. Usually, these latter are checks whose outcome correction requires more skills because, for instance, a high number of false positives is returned and this delicate task is best left in the hands of AlmaLaurea staff.

Gravity and forced data correction

Check outcomes are store in a dedicated table and for each check a line is created for each record that has shown a positive outcome to the check itself. To rank the seriousness of the error and its possible forcing, a system using three colours in a traffic lights is implemented. Cases marked in red

are actual errors that should be corrected. Cases marked in white are suspicious and require a further check. Those marked in green represent forced checks. Forcing means that the record was checked directly by the operator and data were found to be correct.

The first checks performed are domain checks. In SWAL there exist about 70 domain checks that serve to verify the admissibility of the values present in the encoded fields. By way of example, if a record displays the value "V" in the gender field, this is signalled as an error because the only admissible values are "M" and "F". The user may click on the dedicated checks and a grid will be displayed where erroneous data can be corrected. Domain checks, by definition, cannot be forced.

Fig. 2. Example of screenful summing up detected errors

The data correction grid is one of the pivotal elements of the SWAL application. It provides users with data sorting mechanisms and others aimed at performing text replacements in the whole set of data that had a positive outcome after the check.

In the grid, AJAX technologies are utilized for the real-time decoding of encoded fields in the different slots and for the management of pull-down menus containing suggestions during the typing of text in string-type slots connected to domain tables.

From the grid it is also possible to click on the graduate's ID and analyse the relevant administrative data and the list of corrections performed so far on that specific record.

After all domain errors have been corrected, SWAL advises the user to pursue with the administrative check phase that serve to verify the consistency of the graduate record's different fields. For example, if the record contains the name field filled in with 'Francesca' and the gender field filled in as 'M', the system warns that the case is suspicious given that Francesca is a female name. Or, if a record presents with the enrolment_date field = '30/09/2006', graduation_date = '27/05/2008' and prescribed_time_to_graduation = '3 years', the system warns that the record is erroneous because one of the three fields is wrong. Clicking on the check, the user is presented with

a grid where erroneous data can be corrected or where erroneous data can be forced setting the traffic lights on the green colour.

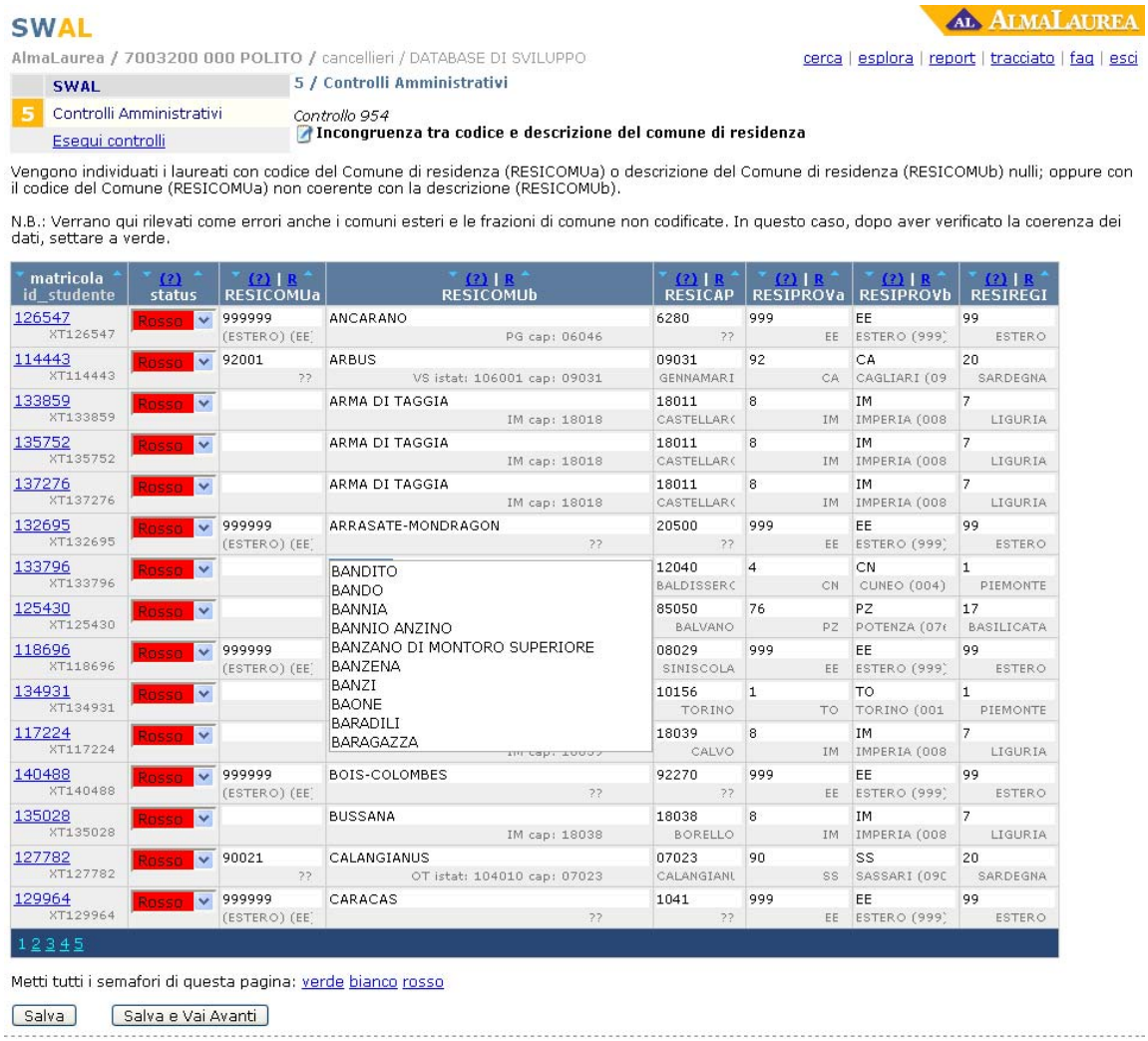

Fig. 3. Example of data correction grid

Every time that the user effects a change and clicks on the save button, the system performs again the whole set of administrative checks. Indeed, many of these are correlated and the correction of data in a field signalled as erroneous may affect the positive outcome of other checks that involve the same field. For efficiency reasons, a cache mechanism has been set that forwards to the database only the queries related to the checks involving modified fields, whereas for the other checks the outcome of the previous operations is considered to be valid.

Once the user has completed the forcing or the correction of all cases signalled by the different controls, he/she is allowed to deliver data. The delivery of data is the last step of the administrative data acquisition procedure which freezes all the data contained in the *officina*/workshop and move them into the final workset.

4.3. Data cleaning – AlmaFlow

After acceptance of administrative data delivered via SWAL, the staff perform a second set of check queries on such data and correct the errors, contacting, if necessary, the faculties' secretary's offices of the different universities that may have access to paper records so as to solve possible inconsistencies. Aside from the inherent value of publishing correct data, these operations are important also in the light of the fact that eventually universities are presented with a report

containing the errors and the corrections performed, and secretary's offices use such reports to update and correct the data stored in their information systems.

After completion of the correction of administrative data, staff users download and upload in AlmaOne the questionnaire data to perform the subsequent matching operation. Questionnaire data originate from the answers given by undergraduates in the online questionnaire and are stored in a remote database (hosted by CINECA). By means of the credentials granted at registration, undergraduates may access at any time the pages of the questionnaire and update the answers. The download of data occurs by means of variable-spacing tabbing-separated text files which are subsequently loaded on AlmaOne. This operation envisages the performance of two steps that are fully automated in some AlmaFlow scripts. The first step is known as import and consists in the loading of data from the text file to a staging table known as "Warehouse". The second consists in the copying of data from the staging table to normalized tables that are built to store questionnaire data.

On completion of data feeding, data undergo a series of preliminary transformations by means of a set of updated queries stored in a dedicated AlmaOne table. Subsequently, a number of basic checks are performed on questionnaire data.

Matching of administrative and questionnaire data

The subsequent matching phase consists of linking each administrative data record to the corresponding questionnaire data record, if existing. To do so, a number of euristics have been defined. These are sets of basic rules for the matching. Given the heterogeneous nature of the information sent to the universities, there are diverse euristics that differ from one another according to the fields in which the matching is performed. Less restrictive euristics are more likely to generate erroneous matchings and are therefore applied only to data not coupled by harsher policies and, in any case, under the operator's control.

The matching based on a given euristic generates a new curriculum ID (ID_CV) and inputs in a table a line with the ID_CV, the administrative record ID and the questionnaire record ID (ID_CV, ID_ADMIN, ID_QUES).

Filling in the questionnaire is not mandatory, therefore not all undergraduates chose to do so. For those for whom questionnaire data are not available, an empty fake questionnaire record is created and matched with administrative data. Users who have not filled in the questionnaire are contacted via the electronic or land mail and are invited to fill it in.

Information merging

Subsequently, the AlmaFlow application moves on to the generation of curriculum data by merging administrative and questionnaire data. Not all administrative and questionnaire data become part of the CV and are published; some of them are stored in AlmaOne and used for statistical purposes only. These include, for example, data related to the assessment of the course that has been just completed.

Once the CV generation phase is finalized, checks are performed on these data and detected errors and inconsistencies are corrected. Corrections are made by the staff based on the CV's entire informational content or by contacting the university's secretary's office, consulting online files and GIS (Geographical Information System) databases.

Fig. 4 - Homepage of AlmaFlow with sequence of operations

4.4. Data publication – CV

When all errors and inconsistencies detected have been corrected or forced and all checks provide a negative outcome, the publication phase begins. This process is fully guided by the AlmaFlow application and includes two stages. In the former, a third series of checks is effected which, if all previous check phases have been performed correctly, should not return any faulty record. The latter stage envisages the CV data extraction and their encapsulation in a text file which is subsequently passed on to CINECA for the updating of the graduate database which can be consulted online. The extraction of data is performed according to a data export protocol stored on an

AlmaOne table which is usually updated before being published. This table, besides the details of the online database fields, such as type an size, memorizes the formula necessary to calculate the field to be exported according to the source fields present in AlmaOne. In the simplest 1-to-1 correspondence cases, the formula field will only display the name of the AlmaOne source field.

At the end of the procedure, the system allows the opportunity to download the text files that have been just generated with the traditional download procedure from a website and transfers the data in the data-warehouse.

4.5. Figures

Currently the database contains 1,148,151 records, of which 1,025,494 have been published and are therefore available for the companies via the online database, broken down by year of graduation with the following figures:

Year of graduation	Overall number of CVs	Published CVs
1996	16,672	16,662
1997	25,092	24,913
1998	33,938	33,547
1999	47,368	45,944
2000	53,820	51,328
2001	67,342	64,297
2002	85,921	77,088
2003	101,641	89,232
2004	142,850	126,406
2005	185,705	160,549
2006	194,929	168,555
2007	192,873	166,973
TOTALE	**1,148,151**	**1,025,494**

Table 1. Graduated present in the database per year of graduation.

At present, about 190,000 graduates are added every year. The latest processing operation collected 77,721 records, of which 65,501 have been published online. The cleaning phase involved the implementation of 1,175 rules on the whole. Out of these, 279 are related to administrative data only, 448 to questionnaire data, 173 to CV data and the remaining 275 are domain checks. Of the 279 administrative checks, 200 are operated by staff personnel only whereas the remaining 79 are preliminarily performed by university staff on the SWAL application.

5. Conclusions

The solution that has been presented enables curriculum data from heterogeneous sources to be collected and, after a number of cleaning and correction phases, it uniforms them and makes them available for employers for recruitment purposes and also for accurate statistical analyses.

The people involved in the data acquisition, validation and correction phases are guided by the available applications (SWAL and AlmaFlow) which also enable operators to flow a pre-set logical pathway in the performance of the different activities.

The solution, besides its high scalability, is totally modular and can therefore be easily adjusted to possible legislation changes in university systems and customized to meet the needs of an individual university. This opens the way to the entry into the AlmaLaurea system also to foreign universities in the hope that a European database can be set up containing accurate and updated information on all graduates of the EU.

6. REFERENCES

Consorzio Interuniversitario AlmaLaurea -2008. *Graduate Profile - 2008 Report* (see also previous works 1994 – 2007), from: http://www.almalaurea.it/eng/universita/profilo/

Consorzio Interuniversitario AlmaLaurea -2008. *Occupational Conditions of Graduates – 2008 Report* (see also previous reports 1996-2007), from: http://www.almalaurea.it/eng/universita/occupazione/

A. Cammelli et al. -2007. *ALMALAUREA: from University courses evaluation to graduates placement.* EUNIS 2007 Conference Proceedings. Abstract online: http://www.eunis.org/events/congresses/eunis2007/CD/pdf/abstracts/p111.pdf

MIUR (Ministry of University) 2001-2007. Database of Education Supply. Available from: http://offf.miur.it

MIUR -2000. DM (Ministerial Decree) of 24 October 2000. Available from: http://www.miur.it/atti/2000/dm001004_01.htm

MIUR -1999. *DM 509/1999, Regolamento recante norme concernenti l'autonomia didattica degli atenei, from: http://www.miur.it/0006Menu_C/0012Docume/0098Normat/2088Regola.htm*

MIUR -2004. *DM 270/2004, Modifiche al regolamento recante norme concernenti l'autonomia didattica degli atenei, approvato con decreto del MURST 3 novembre 1999, n. 509., from: http://www.miur.it/0006Menu_C/0012Docume/0098Normat/4640Modifi_cf2.htm*

R. Kimball, J. Caserta -2004. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley

Stephanie Cammarata -1988. *An Intelligent Information Dictionary for Semantic Manipulation of Relational Databases***.** Proceedings of the International Conference on Extending Database Technology: Advances in Database Technology http://portal.acm.org/citation.cfm?id=649923

Cammarata et al. -1989. *Extending a Relational Database with Deferred Referential Integrity Checking and Intelligent Joins*. ACM SIGMOD Record Volume 18 , Issue 2 http://portal.acm.org/citation.cfm?id=66926.66935