A database for CMU Portugal

Maria Inês Morais
INESC-ID
Instituto Superior Técnico
Lisbon, Portugal
ines.q.morais@tecnico.ulisboa.pt

ABSTRACT

In this day and age, every business, company, or institution can struggle with data storage and organization. With the evolution of technology and the growing need to effectively manage, analyze, and access data, it makes sense to invest in a beneficial data management system. Concerning this topic, the basis for our research problem focuses on creating a database for the Carnegie Mellon Portugal Program (CMU Portugal) and a user interface so users can insert, alter, view, and remove records from the database without being forced to use SQL queries. At the moment, CMU Portugal's data sources consist of various Excel tables and their website. As such, the process of developing a database and user interface for CMU Portugal requires the development of a database model, the profiling, cleaning, and migration of the current CMU Portugal's data into the database, and the elaboration of the user interface. For this process, we will use SQL to build the database, MySQL as the database management system, the Pentaho Data Integration tool to profile, clean, and migrate the data, and HTML, CSS, and Javascript to build the user interface, along with PHP as the server-side scripting language to interact with the database. The implementation of this system is further described in this document. Additionally, the system will be evaluated based on the quality of the data after its migration to the database, the performance of the database, and the user-friendliness of the user interface. Conclusions and considerations for future work are taken at the end.

KEYWORDS

CMU Portugal, Database Design, Data Profiling, Data Cleaning, Data Migration, User interface development

1 INTRODUCTION

More and more, the process of modifying and transforming data is becoming an essential topic in the technology sector. Businesses, companies, and institutions produce large amounts of data that need to be stored and managed in a flexible way. Conversely, databases allow the storage of information quickly and easily and their usage in a wide variety of everyday tasks. At an early stage, organizations may turn to Excel files and similar documents to organize their data. However, this solution is not viable in the long term. This is the case with CMU Portugal, an international partnership program that has grown over the years to include an ever-expanding network of members as well as educational and mobility initiatives. This expansion has made it hard for CMU Portugal to store all of this new data simply and consistently with the help of Excel files alone. CMU Portugal stores its data in multiple Excel files and has, additionally, a website where the data is supplied by other Excel files or handwritten inputs. Given this high number of different sources to respond to different needs, data quality issues arise. Taking this into consideration, our goal is to, first of all, organize CMU Portugal's data in a clear and concise manner through a database model. Secondly, analyze CMU Portugal's data, outline and correct its quality issues. Finally, we aim to create an SQL database exempt from the data problems mentioned before that ultimately integrates the various data sources in order to fulfill both its users' needs and act as an integrated source for the website. Additionally, to achieve this, a database user interface will be developed. This document is structured in six additional parts. After presenting the CMU Portugal program in section 2, a background on its data, sources, and quality issues is presented in section 3. Section 4 presents the related work, section 5 the implementation of the system, section 6 the work's evaluation, and section 7 the conclusion.

2 ABOUT CMU PORTUGAL

CMU Portugal is an international partnership between Carnegie Mellon University (CMU) in Pittsburgh, Pennsylvania, United States, and various Portuguese universities, research institutions, and corporations. This partnership provides a platform for education, research, and innovation, focusing its activities primarily in the area of Information and Communication Technologies (ICT), according to CMU Portugal's annual report [4]. Some of CMU Portugal's initiatives consist of talent development and knowledge creation. Talent development is divided into three programs:

(1) Dual-Degree Ph.D. Programs that provide students with education in two universities, at CMU and, at a Portuguese university. Students finish with two Ph.D. degrees, one from each university.
(2) Affiliated Ph.D. Programs which are fully hosted by a Portuguese university and include a research period at CMU of up to 1 year. Students finish with a Ph.D. degree awarded by the host Portuguese university.
(3) Mobility Programs consist of the Visiting Faculty and Researchers Program and the Visiting Students Program. The first program entails the visit of faculty members and researchers to the corresponding host department at CMU, where collaboration between both parties takes place in research, co-teaching, and other academic activities. The Visiting Students Program enables students to attend courses, participate in research projects, and immerse themselves in the CMU community.

The knowledge creation initiative is made up of the following ventures:

(1) Entrepreneurial Research Initiatives (ERIs) are science, engineering, management, and policy projects that comprise

research teams from two Portuguese universities, one from CMU, and at least one partner company.

(2) Exploratory Research Projects (ERPs) aim to promote new initiatives with high impact potential that encourage Portugal's international competitiveness and innovation capacity in strategic ICT emerging areas of interest to CMU Portugal.

(3) Large-Scale Collaborative Research Projects (LSCRPs) are projects with a higher duration that should involve industrial research and experimental development activities to create new products, services, processes, and systems.

3 BACKGROUND

As previously mentioned, CMU Portugal's data comes from multiple sources, such as various Excel files and handwritten inputs on their website, which leads to data quality issues. CMU Portugal's data and its problems will be further described in this section.

3.1 CMU Portugal Data Overview

CMU Portugal's data is divided into four main topics, the students that participate in the educational programs, the visitors that participate in the mobility programs, faculty members and researchers working under CMU Portugal and the project initiatives fostered by the program.

The students' data consists of their personal data, such as their names, genders, emails, nationalities, and online profiles such as Google Scholar, LinkedIn, and others. In terms of the general data kept about their studies under CMU Portugal, records are kept about the students' research area, degree type (Dual Degree Ph.D. or Affiliated Ph.D.), status and dates of enrollment, expected graduation, and graduation, as well as the number of enrollments. A student's status refers to whether the student is still active, has withdrawn from the program, or has transitioned to "alumni." Since students at CMU Portugal have contact with CMU and another partner Portuguese university, a key element of data kept is the doctorate program(s) the students are enrolled in, as well as the name of the school and university in Portugal and the department and school name at CMU they are affiliated with. Besides that, the names of the advisors at the respective universities are stored, as well as data related to the student's thesis. Finally, the CMU Portugal Program keeps follow-up information on its alumni, such as their current position, location, the position's institution name and type, as well as the starting date. The institution type refers to whether the institution is a university, a government agency, or an industrial institution.

The visitors are students and faculty members coming from Portuguese universities that participate in CMU Portugal's mobility programs. The stored visitors' personal data consists of their names, genders, and emails. The names of the schools and universities where the visitors come from are stored. Additionally, for the student visitors, their highest obtained academic qualification and, for the faculty visitors, their position under the university they come from, are kept. Consequently, data records of the CMU's hosts' names and host departments that receive the visitors are generated. Lastly, the visitors' mobility's start and end dates, as well as the length of stay, are saved.

The faculty and researchers' personal data consists just like the visitors of their names, genders, and emails. The faculty members' and researchers' affiliation(s) stored can include the name of the department, school, and university they currently work under, the research lab they research under, or both. Other important data involves the names of the students they advise and the name and type of project they work on or have worked on.

Projects can be of different types, as we have seen in subsection 2. Important data on the project concerns the researchers involved and their positions in the project. In general, all projects are made up of at least one Principal Investigator (PI) from a Portuguese institution and another PI from the CMU. Moreover, data on the institutions that participate in the project and their roles, e.g. principal contractor, promoter, participant, etc., are also important records. Since projects are usually made up of one Portuguese PI and one PI from CMU, it is also the norm that at least one Portuguese institution and one CMU department are involved in a project.

3.2 CMU Portugal's Data Sources

The majority of CMU Portugal's data comes from various Excel files. Overall, four Excel files with a total of six tables exist. An overview of these tables is provided in table 1. The columns of each of these tables correspond to the stored data of students, alumni, faculty and researchers, and visitors described in the previous section 3.1. Besides that, the tables in these files typically have a large number of columns, some of which correspond to redundant data.

Table 1: Description of the source Excel tables

Table Name	File Type	Description	Nr. of Columns	Nr. of Rows
Ph.D. Students	Excel	Data on Doctorate Students	47	175
AlumniPositions	Excel	Data on alumni work positions	19	85
Faculty	Excel	Data on researchers and faculty members	44	469
Visiting Faculty	Excel	Data on faculty participants in mobility programs	31	88
Visiting Students	Excel	Data on students participants in mobility programs	31	88
Projects	Excel	Data on project initiatives	14	89

Additionally, CMU Portugal has a website that is developed in WordPress and, as previously stated, is fed by Excel files and handwritten inputs. The Excel files that feed the website are formatted in such a way that only the necessary data for the website is retained from the original Excel files. In the back end of WordPress, this data is stored in WordPress' own centralized SQL database, which is developed with MySQL as its database management system [3]. Among the data kept in WordPress's centralized database are [10]:

- Posts, pages, and other content
- Organizational information such as categories and tags
- User data and comments
- Site-wide settings
- Plugin and theme-related data

With this in mind, CMU Portugal's data is scattered across some of these tables, making the WordPress database unusable for the intended purpose. However, through a plugin supported by WordPress, it is possible to download the data referring to the students, faculty and researchers, and projects stored on the website.

3.3 Data quality problems

As previously stated the data sources from CMU Portugal have some data quality issues. These issues are outlined below:

(1) Void excel table entries: Some Excel tables have important unfilled entries that must be completed before migrating the data into the database.
(2) Approximate duplicates in the table's columns: This happens mostly due to mistakes in the input of data. One of these cases happens in the "Ph.D. Students" table in the column "Department in CMU", which stores the department in CMU the doctorate students are affiliated with. This occurs, for example, with the CMU department "Human-Computer Interaction Institute" which also appears as "Human-computer interaction Institute", with the difference being the "c" in "Computer" and "i" in "Interaction" not being capitalized.
(3) Duplicate entries in the "Faculty" table: There are duplicate entries for some faculty members and researchers that contain the same name and represent the same person but have distinct data from each other. This typically happens when old and new data are not merged, leaving two rows of the same person with old and new data in each.
(4) Nomenclature Inconsistencies: There are variations in the definitions of several terms and columns, resulting in the occurrence of non-normalized conventions that can cause some confusion. This happens in the "Ph.D.Students" table in the column "Program in Portugal" which stores the study program of the doctorate students during their stay in Portugal. The same program can have different names, for example, the doctorate program "Engenharia Informática e de Computadores" at Instituto Superior Técnico either appears with that name and the abbreviation "EIC" or as "Programa Doutoral em Engenharia Informática e de Computadores" with no abbreviation given. Nevertheless, these two entries are the same program.
(5) Inconsistency between Excel tables: caused mostly by not up-to-date information. These inconsistencies take place mainly when the names of faculty members and researchers in the "Faculty" table are used in other tables. This happens in the "Ph.D.Students" table in the columns of the advisors, in the "Projects" table columns that detail the names of the researchers, and in the "Visiting Faculty" and "Visiting Students" columns presenting the names of the hosts in CMU. The main cases consist of incomplete or inexistent names of faculty members or researchers in one of these tables.
(6) Missing data from the Excel files that is contained in the website: The necessary columns to extract from the website that are not in the Excel files detailed in table 1 are:
- **Student's data**: Introduction text, research topics;
- **Project's data**: Start Date, end date;

The doctorate students' introduction text refers to a brief text on the website's profile of the student that gives insight into the student's motivation and background. Apart from that, some projects have start and end dates documented on the website that are not present in the original Excel files.

4 RELATED WORK

This chapter analyzes the software tools necessary to accomplish the goals described in section 1. First, database management systems necessary to build the database for CMU Portugal are studied. Data profiling and ETL tools to analyze, clean, and migrate the CMU Portugal data are then explored, followed by an examination of tools to build the database user interface.

4.1 Database Management System

A database management system's (DBMS) main objective is to offer a simple and effective method for storing and retrieving large amounts of data from a database [1]. Examples of database management systems are PostgreSQL, Oracle, and MySQL. MySQL was chosen as the tool to use since, it is important that the chosen DBMS is open-source, supported by Windows, and most importantly, supported by WordPress so that the new database can serve the users' needs and the website. As it is stated in section 3.2 the DBMS of WordPress is MySQL. The Oracle Database makes for a bad contender since it is not open-source and has a hard learning curve [13]. PostgreSQL is a good contender against MySQL since it has better scalability. Nonetheless, the CMU Portugal database does not need a lot of scalability at this point and probably not in the near future because it does not yet reach tens of thousands of records.

4.2 Data Analysis and Integration

Data profiling and ETL tools can help with the process of analyzing and understanding the quality problems of data and ease the process of integrating and migrating this data. The chosen tools to analyze and migrate the data are Pentaho Data Integration and its plug-in DataCleaner. The other options for data profiling are software tools like pandas and the IBM InfoSphere Information Analyzer. While pandas is not a data profiling tool per se, it does have some features that can be used for data profiling, nevertheless, it makes for a bad contender since it does not offer the same level of sophistication as the other tools. Products within the IBM InfoSphere, such as the Information Analyzer and the ETL tool IBM DataStage, are not open-source, which makes them a bad contender. In terms of other ETL tools that allow to profile and migrate data, Talend Open Studio is a good option. Talend and Pentaho fulfill almost all the same qualities, with Pentaho being somewhat faster and having a slightly more intuitive GUI than Talend [6] [2] [14]. Concluding, both tools would be suitable for the development of the project, with the main deciding factors coming down to Pentaho being faster and the lack of experience with the Talend tool.

4.3 Database user interface

The database user interface is the component of a database application that enables users to interact with the database, which offers a graphical interface to enter, view, and alter the data in the database. User interfaces can be developed using HTML, CSS, and JavaScript for the front end and application programs for processing requests at the application server, such as Java Server Pages, or scripting languages such as PHP and Python combined with the Django Framework. It was decided to use HTML, CSS, and Javascript to develop the database user interface since every computer today comes pre-installed with a web browser, and HTML is independent of the operating system or browser, making it usable on any computer that has internet access [1]. Paired with CSS and Javascript, the result can be a sophisticated user interface. PHP was chosen as the scripting language to take requests on the server. This is because Java Server Pages and the Django framework are better suited to develop complex web applications that require significant business logic and database connectivity. Since this is not the case, the option of an open-source scripting language with a simpler learning curve that is flexible and scalable made for a good fit [7].

5 IMPLEMENTATION

In this section, we go over the necessary steps to develop the CMU Portugal database and its user interface. We start off with the requirements gathered for the database, followed by the database model. Besides that, the process of profiling, cleaning, and migrating CMU Portugal's data into the database is described. Finally, an outline of the user interface's development and features is presented.

5.1 Requirements Gathering

The requirements gathering for CMU Portugal's database consists of the main searches carried out in CMU Portugal's various data sources. In order to do this, we identified the most important ones so that our integrated database can satisfy all of these demands. These searches are divided into three categories:

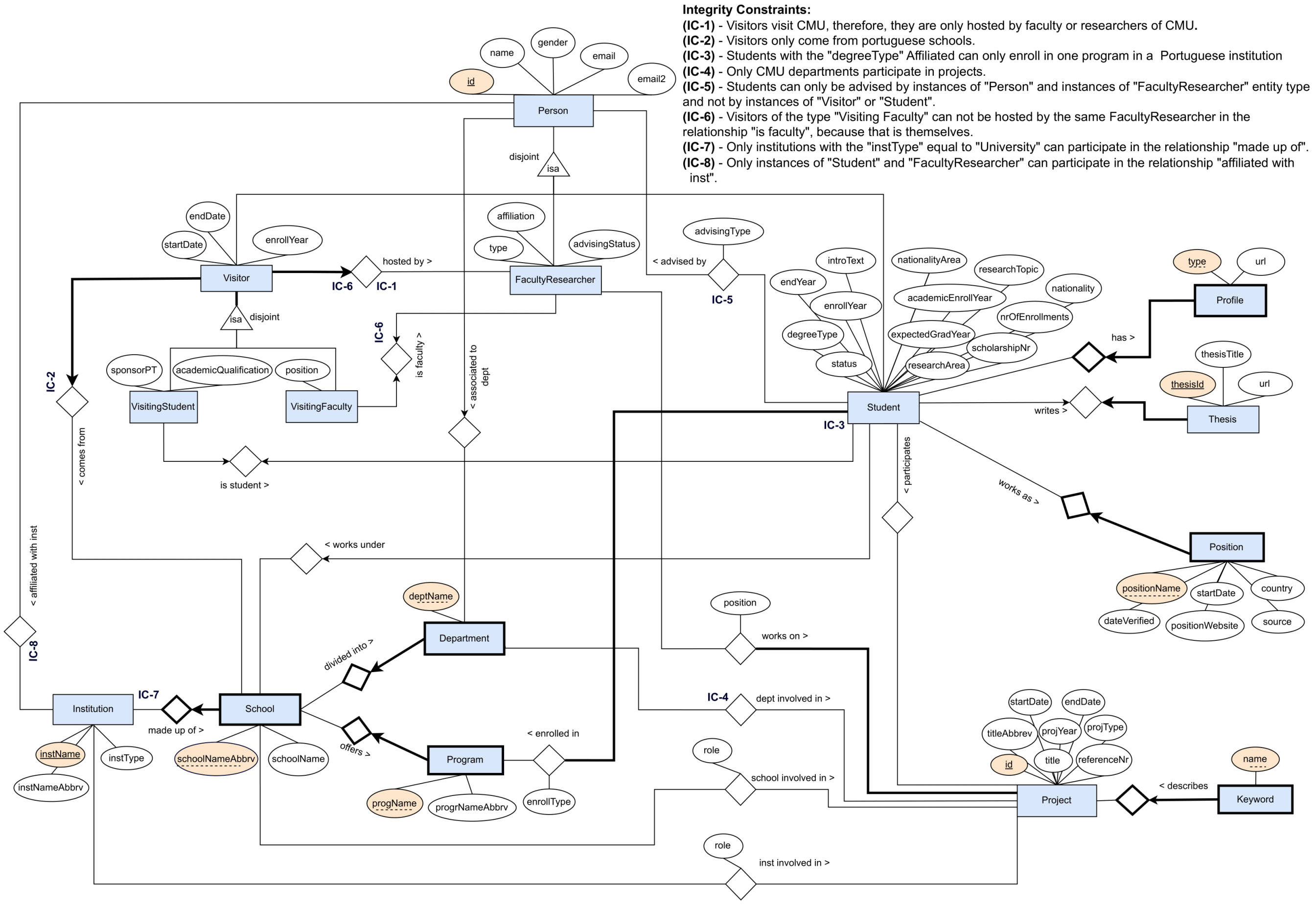
(1) Searches that serve to list the available data so users can view it. These are searches that are likely to be carried out on the user interface.
(2) Searches that are required within CMU Portugal's website functionalities.
(3) Analytical searches the CMU Portugal team needs to perform in order to generate reports.

5.2 Entity-Relationship Model

The developed Entity-Relationship model is presented in fig. 1. A description of this model is provided below. Firstly, we start off by detailing the entities and the attributes associated with them, and then we outline the relationships.

5.2.1 Entities and attributes.

- Person: is a generalization of the entities "Student", "FacultyResearcher" and "Visitor". The attributes of "Person" keep the overall personal data of the lower-level entities. Besides that, the attribute "id" is an internal id created to identify each "Person". In this case, these entities inherit all the attributes and relationships of the higher-level entity they are linked to. Since the constraint on the generalization is not total, instances of "Person" exist. This is for the case of students' supervisors who no longer exist as CMU Portugal's faculty members or researchers.
- Student: an entity that represents students enrolled in CMU Portugal's Dual Degree or Affiliated doctoral programs. This entity's attributes correspond to the student's academic data, and the part of the personal data detailed in section 3.1 that corresponds to their nationalities. The attribute "endYear" refers to the year when a student left the program, either because they graduated or withdrew from it.
- FacultyResearcher: this entity depicts the faculty members and researchers at CMU Portugal that supervise students during their Ph.D. and/or participate in projects included in CMU Portugal's project initiatives. The attribute "type" refers to the type given by CMU Portugal to a faculty member or researcher. The attribute "advisingStatus" designates whether a faculty member is currently supervising a student or not. Finally, the attribute "affiliation" specifies whether the affiliation of the "FacultyResearcher" is to an academic or industrial institution.
- Visitor: portrays students and faculty members of Portuguese universities that participate in CMU Portugal's mobility programs described in section 2.
- Visiting Student: specialization of "Visitor" which represents only the students that participate in CMU Portugal's mobility programs. The attributes specific to these students are their sponsor in Portugal and academic qualification, meaning if they are graduates or master's students.
- Visiting Faculty: specialization of "Visitor" which represents only the faculty members that participate in CMU Portugal's mobility programs. The attribute of "Visiting Faculty" corresponds to their work position.
- Profile: is a weak entity of "Student" that consists of the online profiles of students enrolled in CMU Portugal's educational programs. The attribute "type" refers to the type of the online profile, e.g., "Google Scholar". The attribute "url" corresponds to the link to access this profile.
- Thesis: this entity represents the theses that students of CMU Portugal have written during their doctorate. This entity has an id, to distinguish it from others, the rest of the entity's attributes are the thesis title and the "url" which link can be used to access the thesis.
- Position: a weak entity of "Student" that stands for the work positions of students that have finished their Ph.D. and have become alumni of CMU Portugal. The attributes of this entity are the name of the work position, the starting date, and the country where it is located. Besides that, "dateVerified" saves the date when the work position of the alumni was last checked, "source", the origin of this information, and "positionWebsite", the link where it is possible to find the alumni's position within the company website.
- Institution: this entity represents institutions with whom students, faculty members, researchers, visitors, and project initiatives of CMU Portugal are affiliated. Therefore, these institutions can be of the type "University", "Research Center" or "Company" which is detailed by the attribute "instType". The remaining attributes are the institution's name and abbreviated name.
- School: weak entity of "Institution", represents the schools of a university, such as for instance, "Instituto Superior Técnico" which is a school of "Universidade de Lisboa". The attributes are the name and abbreviation of the school.



Integrity Constraints:
(IC-1) - Visitors visit CMU, therefore, they are only hosted by faculty or researchers of CMU.
(IC-2) - Visitors only come from portuguese schools.
(IC-3) - Students with the "degreeType" Affiliated can only enroll in one program in a Portuguese institution
(IC-4) - Only CMU departments participate in projects.
(IC-5) - Students can only be advised by instances of "Person" and instances of "FacultyResearcher" entity type and not by instances of "Visitor" or "Student".
(IC-6) - Visitors of the type "Visiting Faculty" can not be hosted by the same FacultyResearcher in the relationship "is faculty", because that is themselves.
(IC-7) - Only institutions with the "instType" equal to "University" can participate in the relationship "made up of".
(IC-8) - Only instances of "Student" and "FacultyResearcher" can participate in the relationship "affiliated with inst".

Figure 1: Entity-Relationship Model of CMU Portugal's Database

- Department: weak entity of "School", refers to the departments that exist in a school. The entity's attribute is the name of the department.
- Program: is a weak entity of "School", that describes the Ph.D. programs of a school.
- Project: depicts the project initiatives fostered by CMU Portugal. The entity's attributes consist of the project's year, its start and end dates, as well as the title and the abbreviated title of the project. Besides that, the project's reference number, its type (see section 2), and an id to distinguish between projects are also part of the attributes.
- Keyword: a weak entity of "Project", describes the keywords of a project.

5.2.2 Relationships. We begin by describing the relationships in which the entity "Student" is involved. After that, we move on to the relationships that involve "FacultyResearcher", then "Visitor", then "Person" and finally "Project".

- enrolled in: this relationship between "Student" and "Program" represents the study program(s) the Ph.D. students are enrolled in. Doctorate students may be enrolled in more than one program since students of the degree type "Dual Degree" are enrolled in a program at a Portuguese university and at CMU. On the other hand, students of the degree type "Affiliated" only enroll in a Ph.D. program at a Portuguese university (see section 2). Students must be enrolled in at least one study program. The relationship's attribute "enrollType" defines whether this enrollment is at a Portuguese university or at CMU.
- advised by: while pursuing their doctorate, students have advisors who are faculty members or researchers at CMU Portugal. The relationship "advised by" between "Student" and "Person" depicts this case. The relationship is between "Student" and "Person" and not between "Student" and "FacultyResearcher" since there are older students whose advisors are no longer part of the faculty members and researchers of CMU Portugal and thus can not be represented as instances of "FacultyResearcher". This relationship's attribute "advisingType" refers to whether the supervision between student and advisor took place at a Portuguese university or at CMU.
- writes: this relationship between "Student" and "Thesis" portrays the data about the thesis the Ph.D. student has developed.

- is student: relationship between "Student" and "VisitingStudent", represents the case when "visiting" students that participate in CMU Portugal's mobility programs later enroll and become students of CMU Portugal's doctorate programs. This is a one-to-one relationship, meaning a "visiting" student can only correspond to a Ph.D. student and vice versa since they are the same person.
- participates: is a relationship between "Student" and "Project". This relationship exists because Ph.D. students of CMU Portugal can participate in CMU Portugal's ongoing project initiatives.
- works under: this relationship between "Student" and "School" designates the schools belonging to universities that a former student, i.e., an alumnus, may be working at.
- works on: relationship between "FacultyResearcher" and "Project" that depicts the project initiatives of CMU Portugal on which members of the faculty and researchers work on. The relationship's attribute "position" states the position of the faculty member or researcher under that same project, such as for instance "Principal Investigator" (PI) or "Co-PI".
- is faculty: just as with the "is student" relationship, the "is faculty" relationship represents "visiting" faculty members that participate in CMU Portugal's mobility programs and, later on, become faculty members or researchers at CMU Portugal.
- hosted by: participants of mobility programs, i.e. "Visitors", are hosted by faculty members or researchers at CMU. This relationship between "Visitor" and "FacultyResearcher" portrays the faculty members and researchers that are the hosts to the visitors coming to CMU.
- comes from: is a relationship between "Visitor" and "School" that depicts the schools belonging to Portuguese universities that the participants of mobility programs originate from.
- associated to dept: relationship between "Person" and "Department". This relationship is inherited by "Student", "FacultyResearcher" and "Visitor". Doctorate students are associated with a department at CMU. These students are affiliated with a maximum of one department. Faculty members and researchers of CMU Portugal may also be affiliated with the department at the school and university they work for. "Visitors" are also associated with a CMU department since their participation in a mobility program consists of them being hosted by a CMU department. They are hosted by a maximum of one CMU department.
- affiliated with inst: as with the previous relationship, this relationship between "Person" and "Institution" is inherited by "Student", "FacultyResearcher" and "Visitor". However, only instances of "Student" and "FacultyResearcher" participate in this relationship, as stated in the integrity constraint IC-8. Doctorate students can be hosted by institutions such as research centers, where they conduct their research during their Ph.D. or work for them once they become alumni. The case of "FacultyResearcher" is similar. This relationship describes the affiliation between faculty members and researchers of CMU Portugal who do not work in the academic field but in companies.
- dept involved in: relationship between "Project" and "Department", that depicts the CMU departments that participate in the project initiatives fostered by CMU Portugal.
- school involved in: as in the previous relationship, this relationship between "Project" and "School" portrays the schools that are involved in CMU Portugal's projects. The relationship's attribute "role" describes the role this school plays in the project, for instance, the school may be a "Proponent Institution", a "Participant Institution" or a "Research Unit".
- inst involved in: this relationship between "Project" and "Institution" is similar to the previous ones as it depicts the institutions involved in CMU Portugal's projects. As previously described, "Institution" can represent a university, organization, research center, or company. The relationship's attribute "role" stands for the same as in the relationship "school involved in".

5.3 Data Profiling

From this point on in the project, due to time constraints and in line with the CMU Portugal team's preference, it was agreed to profile, clean, and migrate only the data related to the doctoral students. The developed data profiling process consists of five steps. Firstly, a completeness analysis on the "Ph.D. Students" table and an approximate duplicate detection analysis are performed on this table. The third step consists of detecting approximate duplicates in the columns of the "Ph.D. students" Excel table. After that, a detection of nomenclature issues is performed, and lastly, an approximate duplicate detection analysis is performed between the "Ph.D. students" and "Faculty" tables.

5.3.1 Completeness Analysis. For the completeness analysis, the Pentaho Data Integration plug-in Data Cleaner is used. We select all columns in the ‘Ph.D. students" table that should not be incomplete. After carrying out the analysis, we conclude that of the important fields, there are empty values in the columns that store the student's email, study program, and university at which they are enrolled in Portugal.

5.3.2 Detection of approximate duplicates in the "Ph.D. Students" table. In order to detect duplicates in the "Ph.D. Students" table, the tool Pentaho Data Integration is used, employing a transformation that consists of, firstly, choosing and configuring its two inputs. In this case, the inputs are the "Ph.D. Students" table and an exact copy of that same table. Secondly, we select the columns from each table we want to compare. Given this, we select columns "Id" and "Name" in both steps and rename them to "id1", "name1" and "id2", "name2" to be able to distinguish between them. Following this, the columns coming from both inputs are joined with a step to join rows by using the cartesian product with the following condition: "id1 < id2". This condition is used since there is no need to compare a student's name with its same instance. We then use a step to create a new field called "similarity" which consists of the Levenshtein distance between the fields "name1" and "name2". Finally, we apply a threshold to the similarity and filter the records under this threshold. This transformation did not output any records, meaning there are no duplicate entries in the table "Ph.D. Students".

5.3.3 Approximate duplicate detection in the table's columns of "Ph.D. Students". By detecting approximate duplicates in the "Ph.D. Students" table columns, we aim to discover typing errors and other mistakes. The transformation created to find these mistakes is very similar to the one described in the previous subsection 5.3.2. The only difference is that steps to filter only unique rows were applied before joining the rows. These steps were added since, for example, when comparing CMU study programs, a limited number of doctorate programs appear repeated throughout almost two hundred students. This transformation showed results in the columns that keep track of the students' programs and departments at CMU. The results show that for columns that keep doctorate programs in CMU there are entries where CMU departments were inserted instead. In the case of the columns that store CMU's departments, there are typing mistakes in which some letters are not capitalized correctly and blank spaces at the end of some of the terms inside the columns.

5.3.4 Detection of Nomenclature inconsistencies in the "Ph.D. Students" table's column. In order to detect nomenclature inconsistencies, approximate duplicate detection was not used, because in this case, the variations in the terminologies used are so dissimilar that they cannot be considered approximate duplicates nor detected with similarity measures. Taking this into consideration, another transformation was developed that consists of filtering only unique rows of the columns to analyze. This transformation produced results in the column that keeps the students' study programs in Portugal and the area of the student's nationality. In the column that contains the students' study programs in Portugal, there are at least three programs that are the same but designated with different names, such as, for example "Engenharia Electrotécnica e Computadores" and "Programa Doutoral em Engenharia Electrotécnica e de Computadores". These should both be designated "Engenharia Electrotécnica e Computadores".

5.3.5 Approximate duplicate detection between the tables "Ph.D. Students" and "Faculty". The approximate duplicate detection between the tables "Ph.D. Students" and "Faculty" is performed in order to confirm if the names of the Ph.D. students' advisors are equal to the names of these advisors in the "Faculty" table since the students' advisors are almost always part of the faculty members and researchers of CMU Portugal. In order to find the approximate duplicates, a transformation similar to the one described in the subsection 5.3.2 is performed. The difference is that as inputs, the tables "Ph.D. Students" and "Faculty" are chosen. This transformation is performed for the students' advisors in Portugal and for the ones at CMU. Additionally, another transformation is used to detect the student's advisors in the table "Ph.D. Students" that are not present in the table "Faculty". The results of this transformation indicated fourteen distinct CMU advisors and six distinct advisors from Portuguese universities who had different names in the "Ph.D Students" and "Faculty" tables. Additionally, forty students' advisors' names were not found in the "Faculty" table, some of them due to being written so differently in each table that these were filtered out by the similarity threshold.

5.4 Data Cleaning

The data cleaning process was handled by employing a combination of direct alterations in the Excel file tables and the development of transformations with the Pentaho Data Integration tool.

In the case of the completeness analysis, the results were presented to the CMU Portugal team, which tried to fill in as many of the empty table entries as possible.

Regarding the inconsistencies in the columns that keep the CMU study programs, the CMU departments, and Portuguese study programs, and the student's nationality areas detailed in the subsections 5.3.3 and 5.3.4, direct alterations in the "Ph.D. Students" Excel table were done through Excel's find and replace feature. This solution was chosen since the number of instances requiring manual intervention was low, making it more efficient to handle them directly.

In order to tackle the issue of approximate duplicates between the students' advisors' names in the "Ph.D. Students" and "Faculty" tables, a new transformation was developed. This transformation adds a new column to the "Ph.D. Students" Excel table, and in the rows where the students' advisors' names in the "Ph.D. Students" table do not match with their supposed names in the "Faculty" table, the corresponding name of the students' advisors in the "Faculty" table is inserted in the new column. This extra column enables us to correct these inconsistencies in a more efficient way. However, there is still the issue of the students whose advisors' names are not found in the "Faculty" table. To solve this issue, it was necessary to search through the "Faculty" Excel table to find some of these names since they were written so differently that they were filtered out in the transformations to detect these inconsistencies. The help of the CMU Portugal team was also necessary to ensure the match between these names was done correctly. Besides that, the students' advisors' names that were not present in the "Faculty" table at all were filled in by the CMU Portugal team.

5.5 Data Migration

Once again, the Pentaho Data Integration (PDI) tool was used to migrate the data. Various transformations were developed to migrate the "Ph.D. Students" table's data into several tables of the database. To execute all these necessary transformations in a sequence of steps in a logical order, a job was built with PDI. A job executes the transformations detailed below with a single click.

The first transformation consists of the doctorate students being migrated to the "Person" table. This is done by having the "Ph.D. Students" table as input, selecting the columns with the names, genders, emails, and the newly created internal ids of the students, and mapping these columns to the appropriate fields in the database. After that, the data related to the "Student" table of the database is migrated. In order to achieve this, a similar transformation as the one in the "person" step is performed, in which the difference is that there are two inputs, the "Ph.D. students" table and another Excel table with the students' introduction text and research topics coming from the website. Two transformations were built to migrate the students' advisors to the "Person" database table, one for the advisors at CMU and one for advisors from Portuguese universities. The names and emails of the advisors are selected from the "Ph.D. students" and "Faculty" tables, a new internal sequential id of the type integer is created, and these records are inserted in the "Person" database table. After this, two transformations are needed to migrate the matching advisor's and student's ids into the "advisedBy" database table by selecting these from the "Ph.D. Students" and "Person" database tables and adding a new column with advising type, such as if the advising takes place at a Portuguese university or CMU. A transformation to migrate the CMU schools is performed, which consists of selecting these schools' names, filtering only distinct ones, adding a new column with the Carnegie Mellon University affiliation, and inserting these records into the "School" database table. The transformations corresponding to the migration of the CMU programs and departments into the "Program" and "Department" database tables are very similar to the one that migrates the CMU's schools, with the only difference being that we need to select not only the department names, program names, and program abbreviations but also the schools associated with them. The transformation to migrate Portuguese universities into the "Institution" table consists of selecting only distinct universities' names and adding a new column with the institution type, which is "University" for all rows. The migrations of the Portuguese schools and programs are equal to the transformations to migrate the CMU schools and programs. These transformations migrate the Portuguese schools with their respective university affiliations into the "School" database table and the study programs in Portuguese schools associated with Portuguese schools and universities into the "Program" database table. The transformation to migrate research centers where students do their research is the same as the one to migrate Portuguese universities, with the difference being that the added column contains the term "Research Center" in all rows. In order to migrate Excel data to the "enrolledIn" database table, two transformations are needed, in which the students' ids, study programs, schools, and universities are selected and a new column is added to fill out the enrollment type, "CMU" or "PT". The migrations of the data to the "associatedToDept" and "affiliatedWithInst" database tables consist of selecting the students' ids and their departments at CMU or research host institutions and inserting them into the respective database table. The migration of the student's theses to the "Thesis" database table consists of selecting the theses' titles and links and inserting an internal id of the type integer. The transformation to migrate data to the "writes" database consists of selecting the student's ids and matching theses' ids from the "Ph.D. Students" Excel table and the "Thesis" database tables and inserting them into the "write" database table. Finally, the remaining transformations consist of migrating the students' online profiles into the "Profile" table, which consists of selecting the students' ids and links for a specific profile and adding a new column to insert the profile type, such as "LinkedIn", "GoogleScholar", or other into the "Profile" database.

6 USER INTERFACE

A user interface was created so that the database's users can easily utilize it without needing to resort to SQL queries. With this interface, users can list, create, modify, and remove students. Besides that, it is possible to apply filters to search for, or view only certain students. In the sections below, we provide an outline of the interface's functionalities.

6.1 Search students

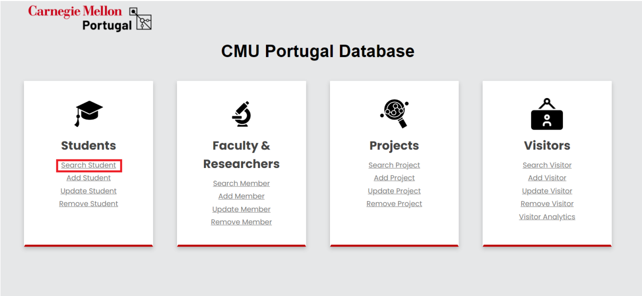


Figure 2: Home page of the CMU Portugal database user interface

Upon entering the user interface, the user sees the contents of the image 2. So far, only the functionalities concerning the students have been implemented. By clicking on "Select Students" the users are presented with a list of all doctorate students, along with their research area, status, degree type, enrollment, end year, and affiliations to Portuguese universities and CMU. All the columns can be filtered with a drop-down menu and a white input box that serves the purpose of writing the value the user wants to filter that column by. The table can also be ordered by the students' names or their enrollment years. By clicking on a student's row, the user is redirected to another page which has that student's profile and contains additional information on that student.

6.2 Add Student

On the home page of the user interface (see fig.2) by clicking on "Add Student" we are redirected to another page that contains a large form that can be used to insert a new student. This form contains forty-three fields, with eighteen of them being mandatory to fill in. Every possible field has a drop-down menu for the user to choose from. Upon submitting, if there are any empty mandatory fields, a pop-up shows up next to the empty field to inform the user that the field needs to be filled in. Once the form is successfully submitted, the student is inserted into the database.

6.3 Update Student

When users make mistakes when inserting students or simply want to update them, the option "Update Student" on the home page can be used. When clicking it, we are redirected to another page with a form equal to the one depicted when the users click on the "Add Student" button on the homepage. There, the users must provide the name of the student they wish to change and are instructed to only fill in the fields they wish to modify. In this form, drop-down menus are also provided for all possible fields so that users can select from the options instead of needing to type in the fields. Once the form is submitted, the selected student's fields are updated in the database.

6.4 Remove Student

By clicking on the "Remove Student" option on the home page of the user interface, the user is allowed to delete a student of their choice from the database. The user is presented with a field to input the student they wish to delete. It is not possible to submit without providing a student's name. After submitting, if a correct student name is inputted, the student is removed from the database.

7 WORK EVALUATION

In this section, we evaluate three key points of this work: quality assurance in the data migration from the source Excel files of CMU Portugal into the database, the performance of this database, and finally, user testing is carried out in the user interface to get their feedback and understand if the platform is intuitive and adequately built.

7.1 Quality Assurance in the Data Migration

Data migration often involves moving large volumes of data from one system to another, for this reason, oversights can result in missing or incomplete data during the transfer. Inaccurate data mapping, which involves matching data fields between the source and target systems, can also result in incorrect data associations, data type mismatches, or loss of data integrity. In order to assure that there are no faults in the migrated data, we use data validation techniques, such as completeness and type correspondence tests [8], which consist of looking at numerous business objects by automating the comparison of these objects coming from the source files and the target database. In order to perform these tests, Pentaho Data Integration is used, in which transformations are created to compare the records in the database to the records in the "Ph.D. Students" Excel table to verify no records were lost or changed.

First, a transformation that consists of a direct comparison between the attributes of the "person" database table and the corresponding columns in the "Ph.D. Students" table is performed. After the transformation runs, the output shows us all the rows that have matched the direct comparison, in which are able to see all 174 students present. This means the data in the "person" database table has been successfully migrated without loss or alterations to the data. A transformation similar to this is performed for the "student" database table, which also successfully outputted all the students. Two different transformations are also performed to confirm the successful data migration into the "enrolledIn" table, one for the enrollments in Portuguese universities and another for enrollments in CMU, which did not point out any mistakes. Two transformations are also needed to test the data in the "advisedBy" database table because students have advisors from Portuguese universities and CMU. No quality issues were found. Lastly, similar transformations as the one performed for the "person" database table are performed for the "writes" and "profile" database tables, with no mistakes identified. In conclusion, with these tests, we are able to verify that no faults appeared in the data due to the data migration process. As such, we guarantee the data quality of the database is up to the same standards as the source data after the profiling and cleaning processes.

7.2 Database Performance

According to Han et al. [9], when benchmarking datasets, the volume of the data, the velocity with which it can be retrieved and generated, and the variety of its diversity define the data. Taking this into account, we wish to ensure that the database responds promptly when users interact with it by performing an analysis regarding the characteristics of the velocity and volume of data in the database. In order to achieve this, we developed a group of queries, identifying relevant workloads, that portrayed the typical behavioral needs the database would need to respond to [9]. In total, we chose fifteen queries, with the first five being queries used in the user interface and with the first three retrieving the highest number of records. The following eight queries represent analytical queries applied to the students in the database. These queries may not return a large number of records, but they all involve retrieving the count of a specific group of students grouped by, for example, their degree type, enrollment year, status, study program, and more, or a combination of various of these attributes. The last three queries are queries related to the functionalities of CMU Portugal's website. After selecting the group of queries to test, we ran each of them three times and got the average time in seconds for each of these queries to execute. We also calculated the average response time of the database based on this group of queries, which is 0.00119706 seconds for an average of 343 records. Taking all of this into consideration, the results obtained show us the database response time for the average query is adequate, being in the millisecond range. This ensures that the users won't be kept waiting or have slow interactions with the user interface, which is the main goal.

7.3 User testing

In order to evaluate the newly developed user interface, we decided to have users test our system. In order to achieve this, we developed a set of four tasks the users should follow that cover all of the implemented functionalities of the user interface. The first task involves exploring the "Search Student" functionality, the second task the "Add Student" functionality, the third the "Update Student" functionality, and the fourth the "Remove Student functionality. After that, we asked the users to fill out a two-part questionnaire, in which the first part consists of their demographic data. The second part consists of users' ratings of the user interface based on a set of heuristics and the detailing of issues in the platform, along with their rating of these issues on the severity scale [11]. The users were handed a description of these heuristics and the severity scale to answer the questionnaire. Five users, who had never before seen or interacted with the user interface, participated in our study.

The heuristics chosen are based on the article by Dowding and Merrill [5]. From the ten heuristics presented in the article, we chose six that we considered to be the most applicable in the context of our user interface, which are the following:

(1) Visibility of system status: The system should always update the user on what is happening by providing suitable feedback in a timely manner.
(2) Match between system and the real world: The system should speak the user's language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms.
(3) User control and freedom: Users should be free to select and sequence tasks when appropriate, rather than having the system do this for them. Users will need a clearly marked exit

to leave the unwanted state without having to go through an extended dialogue.

(4) Recognition rather than recall: The user should not have to remember information from one part of the dialogue to the next.
(5) Consistency and standards: Users should not have to wonder whether different words, situations, or actions mean the same thing.
(6) Aesthetic and minimalist design:Information that is unnecessary or rarely used shouldn't be included in dialogues.

The severity scale consists of a scale going from one to four that rates whether an issue found in the user interface needs to be addressed or not. This scale can be described as follows:

- **0**: I completely disagree that this is a usability issue.
- **1**: Cosmetic problem needs to be rectified only if more time is available to complete the project.
- **2**: Minor usability problem, fixing this should be given low priority.
- **3**: Major usability problem that is important to fix, should be given high priority.
- **4**: Usability catastrophe, imperative to fix before the product is launched.

When designing the tasks for the users to perform, they were designed to be simple tasks, and according to Nielsen Norman Group [12] simple tasks should take about one minute for users to complete. Analyzing the average time taken to complete each task, which was 01.10 minutes for the first task, 01.59 minutes for the second task, 01.26 minutes for the third task, and 00.21 minutes for the fourth task, we observe that the averages are a little over the one-minute mark. However, while conducting these tasks, the users had to read at the same time the guide provided to them, which contained a description of each task. Besides that, the second task had the highest average time, since it involved filling in a form with around nineteen mandatory fields, which context the users did not fully understand. Taking all of this into consideration, we deem that the average duration taken by the users to complete each task is adequate, which implies that our user interface is simple and easy to use.

After this, we analyzed the ratings the users gave the user interface on a scale of one to five, with one being fully disagreeing and five fully agreeing, based on the provided heuristics. In terms of ratings, the first heuristic received an average of 4.8 on the scale, the second 5, the third 3.6, the fourth 4.8, and the fifth and sixth 5. It is possible to analyze the user interface is in accordance with all heuristics except the "User control and freedom" heuristic. When comparing these results to the feedback from the users' issues with the platform, we can understand why. Three users reported the need for a button to go back other than the one provided by the browser or a button that would take them to the home page on all screens. Of these users, all of them rated this issue a two on the severity scale. This means this issue is of low priority, being a minor usability issue. In conclusion, although the issue identified by the users should be fixed, it is not a critical problem that impairs the use of the user interface. Besides that, the requirements of the remainder of the heuristics were met, making this user interface an intuitive and user-friendly platform.

8 CONCLUSION

This paper presented an overview of the development and implementation of a database for CMU Portugal and a user interface so users can list, create, modify, and delete records in this database without resorting to SQL queries. The problem definition was outlined, highlighting the need for an efficient and reliable database to manage CMU Portugal's data, that integrated its various data sources. Overall, the implementation of the database and its user interface addressed the problem at hand and provided a favorable and user-friendly solution for managing CMU Portugal's data. Through completeness and type correspondence tests to analyze the data quality in the migration of the data, a performance evaluation of the database through measuring the duration and retrieved records of queries, and user testing, the developed work demonstrated its effectiveness in meeting the needs of CMU Portugal and its users. The paper highlighted the importance of data quality, database performance, and an intuitive user interface to ensure a reliable and efficient system.

ACKNOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology), under project UIDB/50021/2020, and through the "Carnegie Mellon Portugal Program".

REFERENCES

[1] Henry Kortha Abraham Silberschatz and S. Sudarshan. 2010. *Database System Concepts.* McGraw-Hill.

[2] Md. Badiuzzaman Biplob, Galib Ahasan Sheraji, and Shahidul Islam Khan. 2018. Comparison of Different Extraction Transformation and Loading Tools for Data Warehousing. In *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET).* 262–267. https://doi.org/10.1109/ICISET.2018.8745574

[3] David Damstra Brad Williams and Hal Stern. 2013. *Professional WordPress: Design and Development.* Wrox.

[4] CMUPortugal. 2019. CMU Portugal Annual Report 2018/2019. https://www.cmuportugal.org/wp-content/uploads/2020/09/Relatorio_CMU.pdf

[5] Dawn Dowding and Jacqueline A Merrill. 2018. The development of heuristics for evaluation of dashboard visualizations. *Applied clinical informatics* 9, 03 (2018), 511–518.

[6] Sílvia Castro Fernandes. 2019. Integrating Approximate Duplicate Detection into Pentaho Data Integration.

[7] Devndra Ghimire. 2020. Comparative study on Python web frameworks: Flask and Django. (2020).

[8] Klaus Haller. 2009. Towards the Industrialization of Data Migration: Concepts and Patterns for Standard Software Implementation Projects. In *International Conference on Advanced Information Systems Engineering.*

[9] Rui Han, Lizy Kurian John, and Jianfeng Zhan. 2018. Benchmarking Big Data Systems: A Review. *IEEE Transactions on Services Computing* 11, 3 (2018), 580–597. https://doi.org/10.1109/TSC.2017.2730882

[10] kinsta. [n.d.]. A Beginner's Guide to WordPress Database: What It Is and How to Access It. https://kinsta.com/knowledgebase/wordpress-database/ visited on 2023-04-24.

[11] Daniel Gonçalves Manuel J. Fonseca, Pedro Campos. 2017. *Introdução ao Design de Interfaces.* FCA.

[12] Nielsen Norman Group. [n.d.]. Powers of 10: Time Scales in User Experience. https://www.nngroup.com/articles/powers-of-10-time-scales-in-ux/?fbclid=IwAR0qsOSKooC3wN98YY1RUL5p2Qww_DviAyBeOsdy1lwuouwdBidaFDv7-4w visited on 2023-05-21.

[13] Oracle. [n.d.]. Introduction to Oracle Database. https://docs.oracle.com/en/database/oracle/oracle-database/19/cncpt/introduction-to-oracle-database.html visited on 2023-04-28.

[14] Amanpartap Singh and Jaiteg Singh. 2018. A comparative Review of Extraction, Transformation and Loading Tools. (06 2018).