Type of the Paper (Article, Review, Communication, etc.)

Using the R Language to Manage and Show Statistical Information in the Cloud

Pau Fonseca i Casas [1,*] and Raül Tormos [2]

[1] Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, CA 80034; pau@fib.upc.edu
[2] Centre d'Estudis d'Opinió, Barcelona, CA 08009; rtormos.ceo@gencat.cat
* Correspondence: pau@fib.upc.edu; Tel.: +34-934-017-732

Abstract: We present a methodology to enable users to interact with statistical information owned by an institution and stored in a cloud infrastructure. Mainly based on R, this approach was developed following the open-data philosophy. Also, since we use R, the implementation is mainly based on open-source software. R gives several advantages from the point of view of data management and acquisition, as it becomes a common framework that can be used to structure the processes involved in any statistical operation. This simplifies the access to the data and enable to use all the power of R in the cloud information. This methodology was applied successfully to develop a tool to manage the data of the Centre d'Estudis d'Opinió, but it can be applied to other institutions to enable open access to its data. The infrastructure also was deployed to a cloud infrastructure, to assure the scalability and a 24/7 access.

Keywords: R; Open data; API; Statistics; DSS; Web service.

1 Introduction

The primary goal of the project is to develop a methodology that leads to the implementation of a tool, which currently is on production stage, to analyze statistical information online. To achieve this is needed first a mechanism to manage the large amount of data generated by the surveys and the studies, ensuring that the information remains safe and that the analysts can work with it. Second, a mechanism is required to define what information can be published on the web and what information is not ready to be published (e.g., information that must be anonymized). Finally, a mechanism is required to allow mass media, other research institutions, and the general public to work with the data to obtain new information. To solve these problems, a methodology was defined with the aim of simplifying the interaction with the data of all the actors involved. This methodology was successfully applied on the development of a tool named UPCEO.

This project pursues the idea of open data. The concept of open data to everyone is not new. It was established with the formation of the World Data Center system (WDC) during the International Geophysical Year in 1957 – 1958 [1]. In the beginning, the WDC had centers in the United States, Europe, the Soviet Union and Japan, now it includes 52 centers in 12 countries. The Science Ministers of the Organization for Economic Co-operation and Development (OECD) signed a declaration stating that all the information created or found by the public must be freely available [2]. In that sense, on [3] is analyzed the trend of web, where clearly the semantic representation and access of the information will face the future of access to the information. According to this direction, certain legal tools, such as Open Data Commons [4] came into existence to simplify the use of Open Data over the Internet. In that sense, several tools exist that allow the final user to access information, such as the system in [5], a website devoted to the representation of information on a map, or the Socrata® system [6], a system that supports some interesting applications, such as Data.gov [7] that has the primary mission "*.. to improve access to Federal data and expand creative use of those data beyond the walls of government by encouraging innovative ideas (e.g., web applications).*"

Not only exist several websites and tools to access information but also several applications that allow the reuse and sharing of code related to the access of public information, such as [8] or [9]. The

next step is to allow users without technical knowledge to access the information and perform easy tasks with it. To do this, the user must be able to execute tasks on a remote server that stores both remote information and certain statistical functions.

The possibility to allow end-users to execute certain statistical functions to obtain new information from the data were described by [10]. Several different tools exist to show information over the web and allow the execution of statistical functions by the end users, e.g., the NESSTAR system [11]. In parallel with these proprietary solutions, several efforts are focused to develop APIs to access statistical information. As an example, Data.org is preparing an API that allows users to interact with the system data to build their own applications and mash-ups; the [12] has also implemented an API to interact with its data. However, the question of how to develop and use these APIs remains. Every infrastructure that develops this type of solution implements a new API, and the developers must be able to address all of them.

Another problem is related to the data preparation; several alternatives exist to define the surveys, e.g., [13] or [14]. These tools allow the user to export the data to various formats to perform posterior analyses (a well-known format is the Triple-S, an XML for survey software that enables the user to import and export surveys between different software). The main issue with this approach is that manual operations are required to process the data. In our proposed approximation, once the surveys are completed, they can easily be uploaded in the system, and all of the answers can be related directly to the historical representation of each of the proposed questions. Finally, other interesting concern is related with the standardization for the cloud infrastructures [15]. In our approach we solve this issue proposing the use of R language as the glue to work with statistical information, despite of the cloud infrastructure used.

2 The proposed solution

Statistical institutes that wish to open his data have to deal with intrinsic complexity of their data structures. To arrange these, we could define an ontology as proposed on [16] or [17]. On this paper we propose a different approach, defining a methodology based on the R language see [18] or [19] that simplifies the CRUD (create, read, update and delete) operations that can be performed over the data. To access to this data, we need to face two problems: the format, and the flow. The format is a key factor, since it is needed to define a framework that assures that we can access the data always, see [20] for an example in the judicial area. This can be solved through the use of R as a translation element to our base known format. Also, to be capable to interact with the data, it is necessary to define a flow for the statistical studies that a statistical institution wants to publish. To do so, it is first necessary to categorize the data that we own in the system. We have the surveys that are the elements that lead to obtaining information from the representative sample of the population of study. These surveys must also be managed by the system. In our proposal, they are represented by an initial matrix of data, containing the questions (and the answers to these questions). Because a survey can be related with other surveys (to obtain information over time), it is necessary to define a superstructure to relate the various initial matrices between them at two levels: at the matrix level, and at the table-field level.

Additionally, often some parts of the data obtained from the survey cannot be published (due to anonymity concerns), hence some transformations to the data must be performed to assure the perfect anonymity of the registers. After this is performed, several versions of a study can be published, for example, to correct errors detected in the data. The public must have access only to those matrices of data that pass the necessary quality control, and the other matrices are stored on the system as working matrices but are not accessible to the general public. The life flow of a study is shown in Figure 1, and the structure of our study proposal is shown in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2018 doi:10.20944/preprints201810.0245.v1

Peer-reviewed version available at Technologies 2018, 6, 113; doi:10.3390/technologies6040113

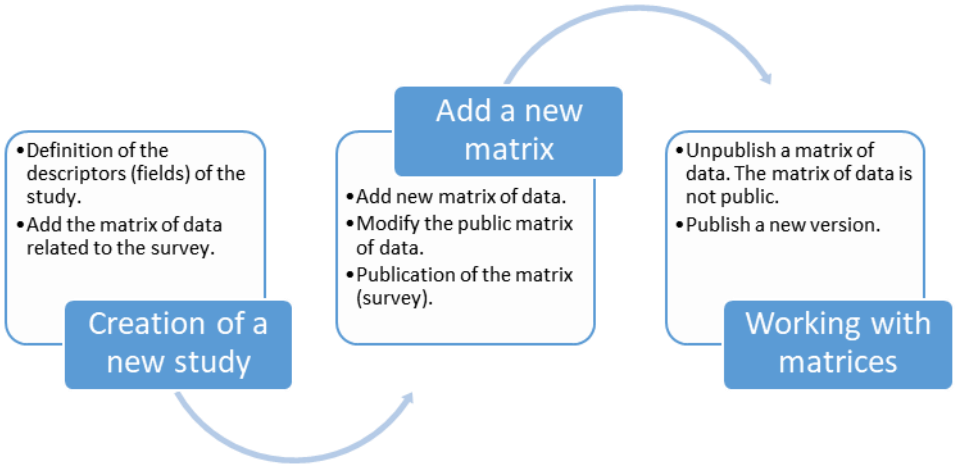


Figure 1. Life cycle of a study.

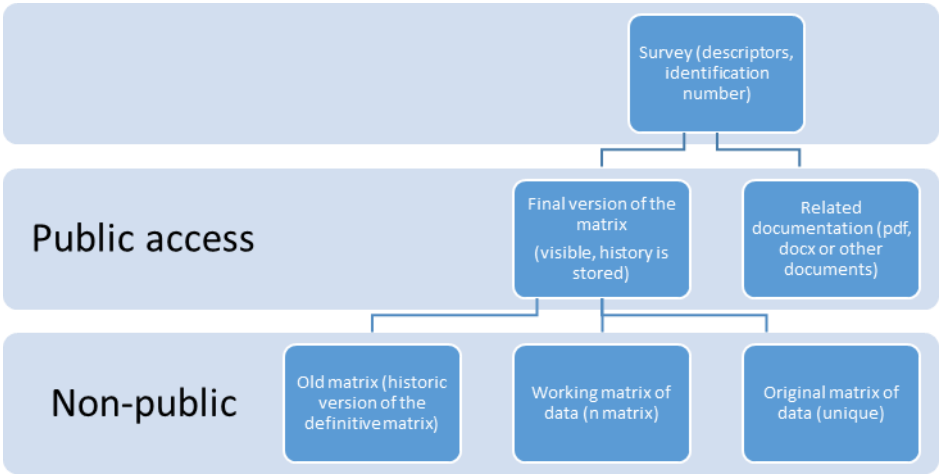


Figure 2. Statistical study structure.

Every study has descriptors to identify the nature of the study and an identification number. For each one of the studies, at least one matrix representing the survey exists. All of the versions obtained from this work are stored in the study structure. Usually, this implies modifying the matrix structures or adding new information. For that, a working matrix exists, representing the last up-to-date matrix related to the studies. The matrix final version is the matrix that the users can operate using R. To manage these matrixes different roles must be defined. Table 1 presents the roles we propose. Each one of these roles has different privileges in the final application. An *analyst* can add new studies, add new matrices to the system, and modify *working* matrices, whereas an *external* user can only perform the statistical operations allowed by the system with the *definitive* matrix.

Table 1. System roles.

Role	Description
Administrator:	Controls the access to the system and defines the roles of the other users.
Analyst:	Manages the information related to the studies (matrix, documentation, etc.)
External:	Can access the system to perform specific operations.

To manage the matrices of data and allow a modification of these data over a cloud infrastructure, worldwide organizations are developing approaches to share statistical information

over the Web using different API. From our point of view, this is not enough because of the inherent complexity of statistical data. Also, this approach would require continuous modifications on the API functions to accommodate them to the new users and institutions requirements that use the data. In our approach, a statistical language is used to provide a common mechanism to access all the information. The data contained in the proposed platform can be published over the internet using the statistical language itself. The result is that the user can interact with the system using the full power of the R language, with no need to define new functions through the API to interact with the data.

2.1 *Beyond the API, using the R language*

We select the R language [18] due to its power and because it is a widely accepted language in the statistical community. R is a free software for statistical computing and graphics; see [21] or [22].

This approach is opposed to the one followed by API development. In our approach, the system allows an authorized user, or program, to access the data and obtain, using R syntax, all the data and information desired. The concern is related not with the implementation of new APIs or protocols to allow access to specific statistical information or data but with limiting the amount of information that can be obtained over the web. This implies limiting the R operations that can be implemented on the server. Fortunately, this configuration can be accomplished through the RServe package [23] [24], which allows the user to define what instructions can be used over the web.

The power of R does not rely only on strong statistical and graphical facilities but also on versatility. Any element of the research community can improve the system by adding new modules to perform statistical operations. One of the packages we need for our approach is RServe. R usually works in standalone applications, and to connect the different services to R, the RServe package must be used. R-Serve can be executed from a command. RServe is a TCP/IP server that allows other programs to use the R facilities from various languages without the need to initialize R or link to the R library [25]. Each connection has a separate workspace and working directory, which is an essential feature for this project. The sequences to start using the service are (i) start the R console, (ii) on the console, load the RServe library.

For most users, the default configuration is satisfactory; however, for this project, RServe must be configured to coordinate the different elements that comprise the system. RServe usually works with several default parameters that can be modified in the *config* file, see [24].

2.2 *Using R on the statistical study lifecycle*

Three main areas must be covered (see Figure 1), the questionnaire management, the management of the matrices related to the study, and the operations management that can be applied to the public study matrices. In each one of these three areas, we propose to use R language as a basic element to simplify the interaction.

To prepare a new questionnaire, first, the questions must be defined. This is not an easy task because of the diversity of questions that can appear, in a single questionnaire, and also because the various surveys (of a barometer trend) must consistently be related to each other to make it possible to obtain accurate conclusions over time. There are various alternatives to design surveys, e.g., [13], or [14] among others. Using these alternatives, the questions can be defined, and they can be sorted on questionnaires that the respondents must answer. Often, these alternatives can export the data to various formats for posterior analysis (such as Triple-S). In our proposal, the relations between the various questions that compose the questionnaires must also be defined; this information (which can be stored in the database for its posterior use) helps us in the review of the complete history of the questions. The answers to the various questionnaires and the history of changes are also available. For example, if we include a question such as, “What party would you vote for in the next election?” and in a new version of a questionnaire, it changes to “If elections were to be held tomorrow, what party or coalition would you vote for?” we must keep the relation between both questions, indicating that they represent the same underlying concept. This simplifies the statistical use in the operations tool, merging the information to construct, for example, a time series.

In that sense, the present approach simplifies the ulterior data management; however, this implies that the uploading process is not easy because it is necessary to create the structure of relationships of the questions, surveys and answers in the database. Additionally, the matrix files can be large and represented in various formats. In our approach, all the information is transformed to a specific XML file that always has the same structure. This enables the user to work with surveys that have the answers in several formats, such as Excel, SPSS, Minitab or R, among many others. Thanks to the use of an XML base representation for the uploading and management of the data matrixes, it is possible to incorporate tools that access the questions. These questions can be presented to the user in various ways, i.e., *editions*. All of the editions of a question can be related, simplifying the operation of merging surveys. The users can build a new questionnaire, and after the questionnaires are defined in the system, they can be related in a matrix that contains the data obtained from the respondents. The key element is to always retain the relation between the questions, the questionnaires and the answers. Finally, and because we propose to use the R language, the users can execute the operations written in R with the data loaded on the system (a subset of the allowed operations).

In Figure 3, the proposed architecture for the management of the statistical information is presented. In this approach, the relation between all of the various questions is preserved. Additionally, R language will be used as an API to obtain information from the system instead of defining an API.

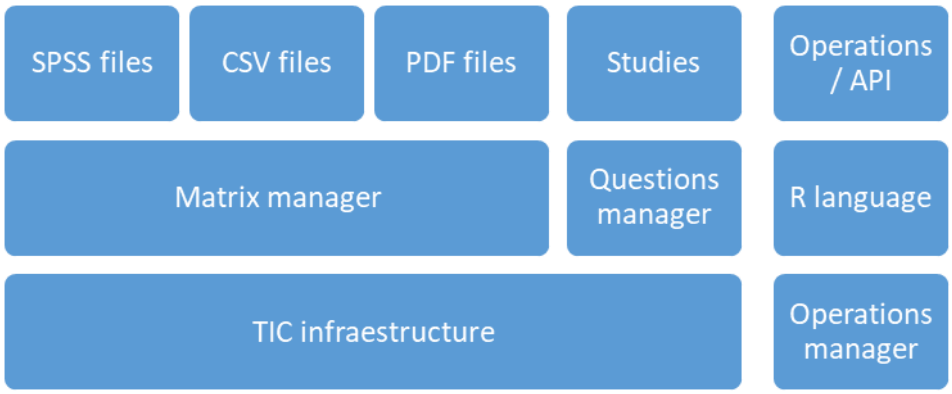


Figure 3. Proposed architecture for the management of statistical information.

3 Case study, the UPCEO application

Three institutions are involved in this real project, the *Centre d'Estudis d'Opinió* (CEO), the InLab FIB and the Centre de *Telecomunicacions i Tecnologies de la Informació* (CTTI). The CEO is the official survey institute of the Generalitat de Catalunya. It handles the government's political surveys, barometers, election studies, and other public opinion polls in Catalonia. As defined in their institutional functions, "It is a tool (the CEO) of the Catalan government aimed at providing a rigorous and quality service to those institutions and individuals interested in the evolution of Catalan public opinion." One of its commitments is to make the information readily accessible to the public.

InLab FIB is an innovation and research lab based in the Barcelona School of Informatics, Universitat Politècnica de Catalunya - Barcelona Tech (UPC) that integrates academic personnel from various UPC departments and its own technical staff to provide solutions to a wide range of demands that involve several areas of expertise. InLab FIB, formerly LCFIB, has more than three decades of experience in developing applications using the latest ICT technologies, collaborating in various research and innovation projects and creating customized solutions for public administrations, industry, large companies and SMEs using agile methodologies.

The *Centre de Telecomunicacions i Tecnologies de la Informació* (CTTI) [26] is an infrastructure that can host all of the services that the various organizations that belong to the *Generalitat de Catalunya* require. This infrastructure is maintained by a licensed private enterprise (now T-Systems). This is convenient for the project because, when the CEO releases a new study, the quantity of resources

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2018 doi:10.20944/preprints201810.0245.v1

Peer-reviewed version available at Technologies 2018, 6, 113; doi:10.3390/technologies6040113

required to supply the punctual demand can be bigger than the resources required in a usual day. Additionally, because CTTI ensures that the system is working 24/7, it can be convenient for the daily work to provide the infrastructure for the CEO database to store all of the information regarding the studies. The CEO primarily manages surveys related to political public opinion. The studies derived from these surveys are published on the CEO website to ensure that the public has knowledge about the studies.

Based on the proposed architecture presented in Figure 3, we implement a system for CEO to simplify the management and use of statistical information over a web. The specific implementation is represented in Figure 4. The system is composed of different layers, each one of which is related to the various services that the system must provide. The web server is based on a WebLogic Oracle® application [27], using Apache Struts [28] [29] and Java as the infrastructure to define the interface of the system and to establish communication with the R system. The main purpose of using R is to implement various operations that deal with data. As an example, we use R to obtain the data from the matrix and the surveys that usually are in the original form of Excel spreadsheets, SPSS files or SAS files; here, R is used as the bridge between all of the various file formats. The R language can be used by users and other applications as an API to communicate with the system to obtain statistical data. In Figure 4, the structure of the system is shown. The entire system is on the CTTI cloud infrastructure. The various files related to the application are stored on an NAS system. The studies are stored in an Oracle database to manage the various files of the system. The R application is installed on the system with the RServe package, defining a set of operations (as an API) and publishing them on the internet using WebLogic platform.

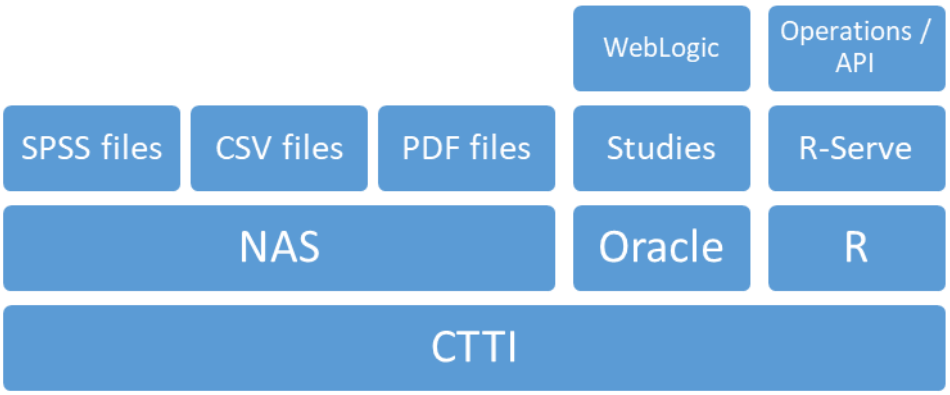


Figure 4. System structure.

From an operations point of view, when a user requests a specific study, he obtains its related documents, mainly *.pdf* files and links to other data related to the survey. With these data, the user can perform various operations (with R), obtaining new data and information. These results can then be exported in CSV file format to be analyzed in more detail using any statistical package. As shown in Figure 4, the matrix is stored in its original form on the NAS, implying that various formats must be stored in the system. In this way, the information generation process can be reproduced exactly as it was by the analyst.

The main file formats that can be used by the CEO analyst are Excel spreadsheets, SPSS *.sav* files and *.csv* files. The various functionalities in the system are:

- **Questionnaire manager** manages the questions related to each one of the different questionnaires of the system; see Figure 5. All of the questions must be related, to allow an analysis over time of the data stored on the database.
- **Matrix manager** manages the information related to the matrix generated by the surveys; see Figure 6.
- **Operations manager** shows the information to the users and other applications (websites) through the R language.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2018 doi:10.20944/preprints201810.0245.v1

Peer-reviewed version available at Technologies 2018, 6, 113; doi:10.3390/technologies6040113

The application can be accessed at http://ceo.gencat.cat/ca/inici/. The website (in Catalan language) gives access to the operations in "*Banc de dades del BOP*," located at the bottom of the page.

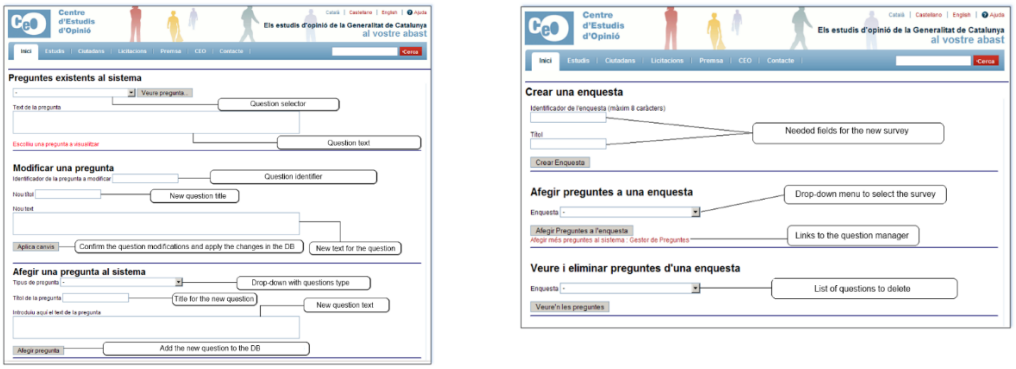


Figure 5. The process of creating a new question (left) or a new survey (right) is integrated into the application, simplifying the process of reuse and relating the questions of all the questionnaires that exist in the system, as is proposed by our approach.

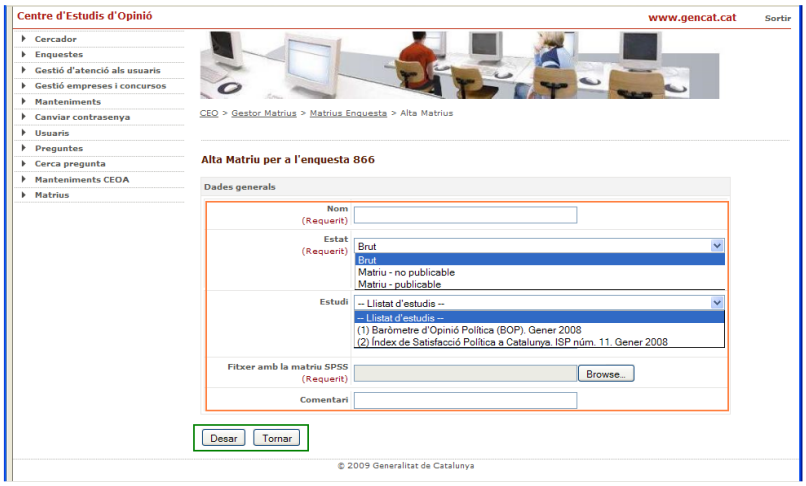


Figure 6. Uploading a new matrix containing the data of a survey to the system.

4 Implementation and calibration of the application

The entire application resides as a cloud solution supported by the *Generalitat de Catalunya*, hosted by the *Centre de Telecomunicacions i Tecnologies de la Informació* (CTTI). In this cloud solution, the options to work and to modify the code uploaded are limited, as is explained in section 3. Because of the complexity of the structure and the required security concerns, a test infrastructure was implemented to test the R operations. The test infrastructure is composed of a server and a client. On the server side, a machine acts as a Web server (using IBM WebLogic), hosting the MySQL database, storing the data on the NAS (Network Attached Storage) and executing R-Serve. On the client side, a java program (named JGUIforR; see Figure 7) is used to define the GUI and the R code needed to execute the operations and manage the matrices.

The client application must first be connected with the server side. The IP of the R server instance we want to use is defined. In this case, the application is connecting with a server that is executed on the same machine as the JGUIforR.

Once this is completed, the connection with the server is established using the File menu. Two options are available. **RComand** implies that the user is working with a local instance of R. In that case, it is not necessary to define the IP. **RComandTCP** implies that the user is working with a remote instance of R; in that case, the IP of the remote server must be defined. If the connection is established

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2018 doi:10.20944/preprints201810.0245.v1

Peer-reviewed version available at Technologies 2018, 6, 113; doi:10.3390/technologies6040113

without error, a message appears in the **R Comands** window showing the version of the R engine used on the server side.

To start working, a dataset must be selected, in this case, an SPSS® dataset. Opening a new dataset is as easy as going to the File menu and selecting a new **Matrix** of data.

Once the matrix is loaded, a message is shown to the user in the **R Comands** area, as shown in Figure 7. At this point, all the operations are active, and the user can start working with the matrix.

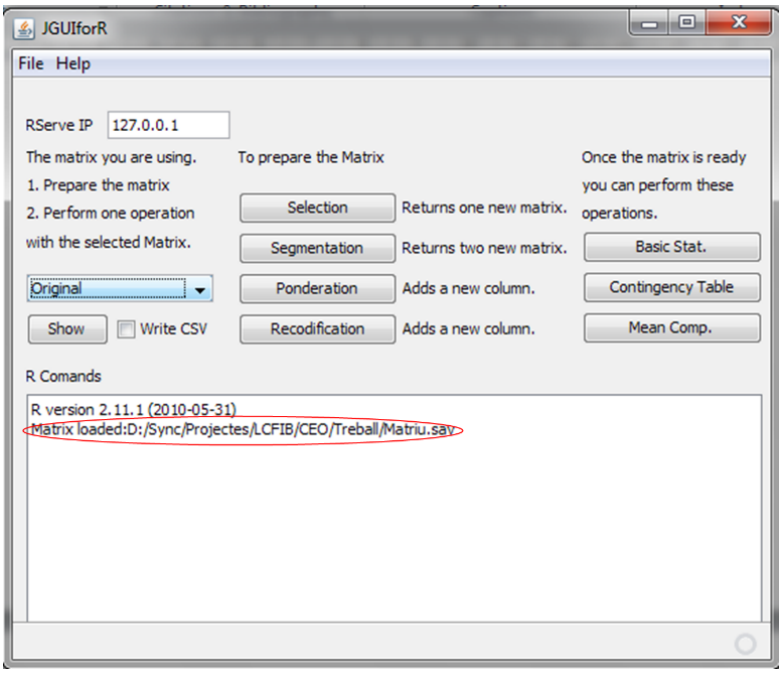
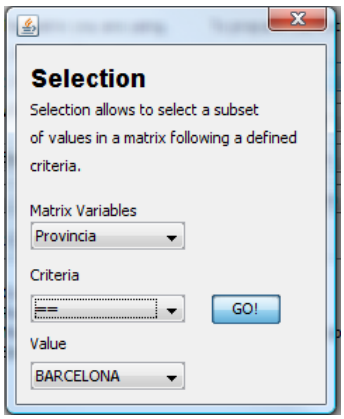
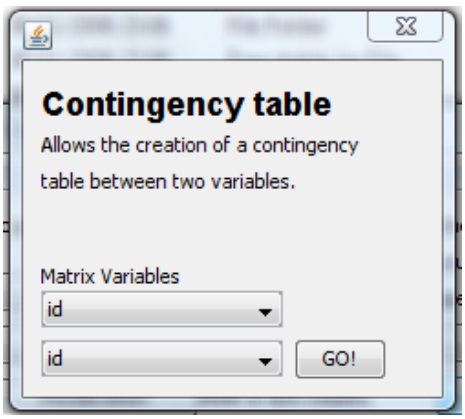


Figure 7. Matrix successfully loaded. All of the options are now activated, and the user can start working with the matrix.

CEO analysts use this software to understand the operations that the system publishes and, to understand, the behavior desired in the final implementation of the client, see Figure 8. As shown in Figure 7, the operations are divided into two main groups. The first includes the preparation of the matrix, selection of a portion of the data of the entire matrix, segmentation of the matrix, weighting of some of the columns of the matrix and recodification. The other operations that can be executed works over this matrix. The results are obtained also following R syntax, as example **Error! R eference source not found.** shows the code that allows us to obtain the contingency table that will be shown to the user as in described on 4.3.3.





```
Selection_BARCELONA <- subset(Original, Provincia=="BARCELONA")

.Table <- xtabs(~<variable_1>+<varialbe_2>, data=<Matriu>)
.Table
totPercents(.Table) # Percentage of Total
.Test <- chisq.test(.Table, correct=FALSE)
.Test
remove(.Test)
remove(.Table)
```

Figure 8. JGUIforR Selection and Contingency Table operations and its R code. R code to select the rows in a matrix that have the value of "BARCELONA" in the variable Provincia (province) is shown (left). In the right is show the GUI used for the contingency table operation (see section 4.3.3).

Next you can see the contingency table R code answer. The answer obtained is a double entrance table, with total percentages and a chi-square test to analyze the independency.

```
[REAL* (318.0, 190.0, 189.0, 192.0, 330.0, 194.0, 195.0, 192.0)]
[VECTOR ([REAL* (0.08876325969681931)], [REAL* (3.0)], [REAL* (0.9931509045238127)], [STRING "Pearson's Chi-squared test"], [STRING ".Original"], [REAL* (318.0, 190.0, 189.0, 192.0, 330.0, 194.0, 195.0, 192.0)], [REAL* (320.04, 189.65333333333334, 189.65333333333334, 189.65333333333334, 327.96, 194.34666666666666, 194.34666666666666, 194.34666666666666)], [REAL* (-0.11403234005394253, 0.025172818498354458, -0.04744108101613011, 0.1704006175273236, 0.11264702552683832, -0.024867008136684843, 0.046864746103752755, -0.16833051661756004)])]
```

4.1 *Deploying the system*

Once operations perform as expected, the system can be deployed on the CTTI infrastructure. This project represents the first deployment of RServe on the CTTI infrastructure, which implies the need to define roles and protocols to ensure 24/7 support. First, the application is deployed on the working server, a machine accessible only to the computers located at the InLab FIB. Once the application passes the tests on these machines, it is deployed at the integration level of the CTTI infrastructure. Here, the application is tested in an environment that has similar security levels and the same software. After the application performs well there, it can be deployed to a preproduction level. Here, the application runs on an exact replica of the final infrastructure, on the same hardware and executing the same software that the application will find in the production environment. At this level, a set of tests are performed, and the application must pass all of them to be deployed to the production level.

At the production level, the application is available for public use. Once the system is deployed, the operations performed by the users never modifies the information stored in the server. The system must also be able to store information regarding the various activities that each of the users do. When an operation is selected, the R syntax is stored in the database. This syntax is not executed immediately on the system; it is only executed when the user requests the results (for example, executes the operations of basic statistics).

4.2 *Testing the system*

To understand how the system performs, it is necessary to test how it behaves during each of the various tasks that must be accomplished in the process of producing and analyzing a new matrix of data. In that sense, the main operations are (i) the process of preparing the surveys and uploading a matrix of data and (ii) the process of operating with these data and returning the information to the user.

4.2.1 Preparing the questionnaires and uploading a matrix of data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2018 doi:10.20944/preprints201810.0245.v1

Peer-reviewed version available at Technologies 2018, 6, 113; doi:10.3390/technologies6040113

As is described in section 3, the questions and the questionnaires must be related to assure obtaining chronological information. This implies that the matrices of data are huge and have a great impact on the database. These incur on increased time to upload a new matrix of data in the database. Several alternatives were analyzed and tested to reduce this upload time. The first alternative tested is the use of SQL commands to include the new answers on the database (creating the relation with the questions and questionnaires of the surveys). This approach was not possible to implement because the file can be larger than the maximum value of bytes accepted by the database to be included in a single operation. The second alternative, which is uploading the files by parts, does not solve the problem because the time needed to upload the complete file grows exponentially from the communication protocol of the server, shown in Figure 9.

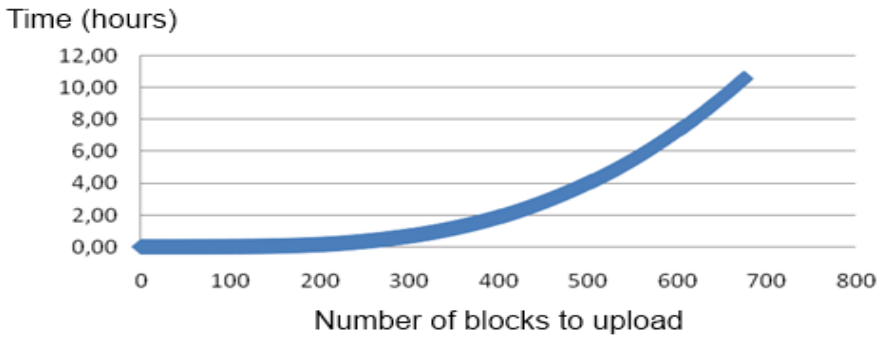


Figure 9. Time required to upload a new matrix of data, depending on its size, using SQL commands.

Other alternatives that offer good performance are SQL LOADER (Moore, 2003) or External Tables (Billington, 2007). However, both techniques, although dramatically improving the time required to upload the matrixes of data (more clearly with large matrixes), are not suitable for use in our cloud infrastructure. The implemented solution was based on the use of CLOB and XMLType (ORACLE-BASE.com, 2012). Test results showed that, to insert the XML file with the XmlType format in the database quickly, it must first be inserted as an object of the type Oracle CLOB and then reinserted into another object of the type XMLType.

With this solution, the time required to upload a matrix of data is less than two minutes for a normally sized matrix and approximately five minutes for the large matrixes. With this approach, not only is the matrix uploaded, but the relationships for its posterior calculus are also created and stored.

4.2.2 The process of operating with the matrix of data and returning the information to the user.

The tool must operate at least as efficiently as if the user downloads the matrix of data and uses R on a desktop computer.

All of the operations are stored in the database using the R syntax. Thus, when a user starts the operations, the system stores the R code in the database, but as we said, it does not execute the code. The code is only executed when the user requests results. This is because the system has a delay in the establishment of the connections between the client and RServe. This delay can be of a minute in some cases. Using this approach, the delay is minimized. After the connections are established, R performs well and returns the new data very quickly. In all of the tests, the time required to obtain new values, after the initial connection was established, depended only on the web latency.

One of the main advantages of the new system is that it does not allow the users to accidentally modify the data, and does not allow mistakes in the statistical operations.

4.3 Examples of use, operations manager

In this section we show an examples of application manager use. First the user must select the initial survey (matrix) to work (see Figure 10).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2018 doi:10.20944/preprints201810.0245.v1

Peer-reviewed version available at Technologies 2018, 6, 113; doi:10.3390/technologies6040113

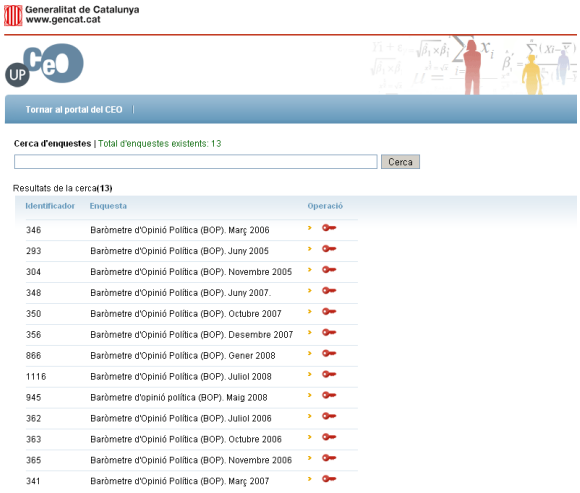


Figure 10. Selection of the matrix.

Once the user selects a study to be analyzed, a new window is shown. In this window the basic information of the matrix is portrayed. A menu on the left side also shows the operations that can be applied to the matrix of data obtained from the survey. Figure 11 presents this window. Because each one of the different operations are implemented using the R language, the implementation of new (and maybe more complex) operations is simplified to solving two problems. The implementation of the operation is programmed using R code, and the output is formatted to make it visible over the web.

Figure 11. Once the matrix is selected, the user can perform some operations to the data. In the figure,, the operations that can be performed over the selected matrix of data, in this case the "Baròmetre d'Opinió Política (BOP) of 2010, are shown on the left.

Figure 12. Working with matrices implies the generation of new matrices to preserve the data on its original state.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2018 doi:10.20944/preprints201810.0245.v1

Peer-reviewed version available at Technologies 2018, 6, 113; doi:10.3390/technologies6040113

Each operations grouped under the label *Preparació de dades* (data preparation), generates a new matrix of data that can be further analyzed (see Figure 12).

The operations that can be performed in this case are:

- Selection (*Selecció*): this operation allows selecting a subset of the matrix according to a specific condition.
- Segmentation *(Segmentació)*: this operation splits the matrix into two different matrices according to a specific condition.
- Recodification *(Recodificació)*: this operation modifies the variables of the matrix according to a rule.

In this first version of the website only one preparation of the matrix can be done, implying that all the preparations are always applied to the "definitive matrix" (see Figure 2). Once one of these operations is performed, a new matrix will be created. The new matrix is presented in the drop-down menu, and can be selected as a data source.

Ponderation operation emphasizes different values of some variables over others. Ponderation is usually needed when the samples used (although are enough to develop the study) are not proportional with the population they represent. To work with ponderation the user must select a check box that appears on the foot of all operations, see Figure 13, Figure 14 or Figure 15. The ponderation values are defined by the analysts on the original matrices (in a column in the original SPSS® matrix), becoming an element that can be used by a non-expert user in a transparent way. As examples of the different implemented operations selection, basic statistics operation, and contingency table are described next.

4.3.1 Selection operation

Selection chooses a subset of the data stored in a matrix, see Figure 13. The process that a user must follow to perform the operation is (i) the user selects the matrix to be used to perform the operation; (ii) the user selects the variable to perform the selection; (iii) the user defines the new values for the new variable; (iv) the system adds the new variable to a new matrix at the drop-down menu. The result of the operation is stored in a new matrix that has the name Selection_<variable>. Now we can perform operations to this newly created matrix.

Figure 13. Selection operation.

4.3.2 Basic statistics operation

We can obtain a basic statistical analysis over an UPCEO matrix. The process is (i) the user chooses the matrix to be used to perform the operation (optional, the default is the *original* matrix); (ii) the user selects the variables on which perform the statistical operation, descriptive statistical

analysis; (iii) the system shows the result of the selected operation in the window over the selected variable, see Figure 14.

Figure 14. Basic statistics results.

4.3.3 Contingency table

This operation allows the user to analyze the relation between two or more variables. The process is (i) the user selects the matrix to be used; (ii) the user selects the variables to be used; (iii) the system shows the tables, the contingency tables over the variables, the percentages values over the total, the percentage by rows and columns, see Figure 15.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2018 doi:10.20944/preprints201810.0245.v1

Peer-reviewed version available at Technologies 2018, 6, 113; doi:10.3390/technologies6040113

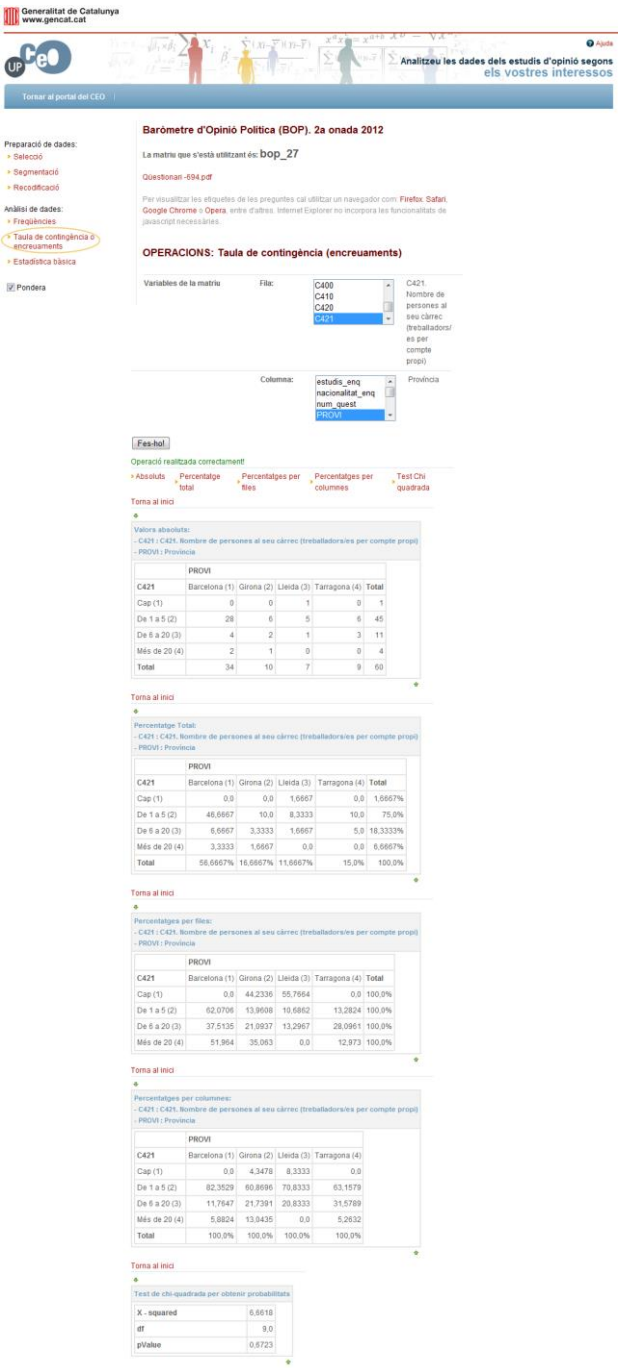


Figure 15. In the "Taula de contingencia (encreuaments)," the user obtains a bivariate analysis of contingency tables for two selected variables. The analysis involves several R commands that completely describe the variables involved in the analysis. The user does not need any knowledge of R or its syntax; the user just selects the two variables to analyze and the system ensures the correctness of the information presented.

5 Concluding remarks

This study develops a novel approach to present statistical information over the web following the open-data philosophy. In this approach, the R statistical package is a key element to manage and display the information, allowing the user to perform a number of statistical operations with the data.

From the point of view of data management, the structure of the surveys, the structure that relates the questionnaires and the questions and the related matrix that contains the data, often follow different formats in an actual environment. This is true even if a single team manages the information because technology changes and the tools used can be diverse, depending on the objectives of the

specific work. This ecosystem of data formats often makes working with the data more difficult. Thus, mechanisms are necessary to translate the information from one format to another. Often, these mechanisms are prone to errors and require the use of tools that are usually not well-known by all of the members of the team. In this approach, R is the bridge between the various formats that are stored in the database and is also the language used to recover and work with the information contained in the system. Thus, the CEO analysts store the information in the system using the format they are familiar with, and the system is able, using R, to work with the data and to formulate new matrices of data that can be used again by the experts applying the common statistical tools.

A cloud solution is implemented to simplify UPCEO management and scalability. The flow of access of the external users depends on several factors, e.g., when a new study is offered to the public. This implies that, at times, the traffic to the site is heavy, an aspect that can become a problem for the servers and site management. The cloud solution proposed stores all the information obtained from the CEO studies, allowing 24/7 access to all the information by all the users, and making it possible, conditional on the user role, the manipulation of the data and the creation of new information and matrices. Working with the data is accomplished using R as a statistical engine; a user can execute queries and obtain new information regarding the matrixes of data related to a survey. Additionally, as all the operations implemented use R syntax, adding new operations is easy and only requires the addition of a new R code and the definition of a new interface. Thus, systems based on this approach are extremely scalable and expandable.

Since all of the access to the statistical information is based on the R language, new websites or applications (such as JGUIforR) can be developed that access the data through the use of R statements. This implies that the application goes further than the definition of an API because it uses a statistical language. The power and extensibility of R ensures that we can obtain all the information needed, and the user must only define the subset (if it is needed) of the R instructions that an external user (application or website) can execute. Currently, researchers from various Catalonian institutions are building their own mash-ups using the application. In the future, more capabilities will be added to the application by adding new R language instructions open to public use. There is an additional goal of open access to the institutions, allowing them to access all the information from the CEO servers and define the queries needed for each application (in the broad sense that an application can be a simple query that can reside in a spreadsheet, or a complete web application with various mash-ups).

Last but not least, a set of operations can be defined as an R script. This definition implies that repetitive operations can be performed with fewer errors and in less time.

6 Bibliography

[1] World Data Center, "World Data System of International Council for Science," 2010. [Online]. Available: http://www.icsu-wds.org/. [Accessed 11 11 2011].

[2] Organisation For Economic Co-operation And Development, "OEOECD Principles and Guidelines for Access to Research Data from Public Funding," 2007.

[3] E. Khan, "Big Data , Natural Language Understanding and Intelligent Agent based Web," in *Proceedings of the 9th International Conference on Computer Engineering and Applications (CEA '15)*, Dubai, United Arab Emirates, 2015.

[4] Open Knowledge Foundation, "Legal tools for Open Data," 2011. [Online]. Available: http://opendatacommons.org/. [Accessed 11 11 2011].

[5] open3, "DataMaps.eu," 2011. [Online]. Available: http://www.datamaps.eu/. [Accessed 11 11 2011].

[6] Socrata, Inc, "Socrata, The Open Data Company," 2011. [Online]. Available: http://www.socrata.com/. [Accessed 11 11 2011].

[7] Federal Government, "Data.gov Empowering People," 2011. [Online]. Available: http://www.data.gov/. [Accessed 11 11 2011].

[8] Code for America Labs, Inc , "Code for America," 2011. [Online]. Available: http://codeforamerica.org/. [Accessed 14 11 2011].

[9] Leipziger Agenda 21, "API.LEIPZIG," 2011. [Online]. Available: http://www.apileipzig.de/. [Accessed 14 11 2011].

[10] B. Sundgren, "Making Statistical Data More Available," in *Workshop on R&D Opportunities in Federal Information Services.*, Virginia, USA., 1997.

[11] T. Assini, "NESSTAR: A Semantic Web Application for Statistical Data and Metadata.," in *WWW2002 Conference.*, Hawai, 2002.

[12] New York State Senate, "NYSenate.gov Application Protocol Interface (API)," 2011. [Online]. Available: http://www.nysenate.gov/developers/api. [Accessed 14 11 2011].

[13] Snap Surveys Ltd, "Online surveys," 2012. [Online]. Available: http://www.snapsurveys.com/. [Accessed 20 10 2012].

[14] University of Ottawa, "Snap Surveys," 2012. [Online]. Available: http://www.ccs.uottawa.ca/webmaster/survey/. [Accessed 20 10 2012].

[15] A. Koschel, S. Hofmann and I. Astrova, "Standardization in Cloud Computing," in *Proceedings of the 8th WSEAS International Conference on Computer Engineering and Applications (CEA '14)*, Tenerife, Spain, 2014.

[16] A.-b. M. Salem and S. Cakula, "Using Ontological Engineering for Developing Web-Based AI Ontology," in *Proceedings of the 6th International Conference on Communications and Information Technology (CIT '12)*, Vouliagmeni Beach, Athens, Greece, 2012.

[17] A. Pomp, A. Paulus, A. Kirmse, V. Kraus and T. Meisen, "Applying Semantics to Reduce the Time to Analytics within Complex Heterogeneous Infrastructures," *Technologies*, vol. 6, no. 3, p. 86, 2018.

[18] J. Adler, R in a Nutshell: A Desktop Quick Reference, O'Reilly Media, 2009.

[19] P. Teetor, R Cookbook, O'Reilly Media, Inc., 2011.

[20] F. Amato, A. Mazzeo, V. Mostato and A. Picariello, "A system for semantic retrieval and long-term preservation of multimedia documents in the e-government domain," *International journal of web and grid services*, vol. 5, no. 4, pp. 323-338, 2009.

[21] F. Murtagh, Correspondence analysis and data coding with Java and R, C. S. a. D. A. Chapman and Hall, Ed., 2008.

[22] M. W. Trosset, An introduction to statistical inference and its applications with R, vol. 81, Chapman and Hall., 2010.

[23] S. Urbanek, "Rserve A Fast Way to Provide R Functionality to Applications," in *Procediings of the third international workshop on 'Distributed Statistical Computing'* , Vienna, 2003.

[24] Rforge.net, "Rserve - Binary R server," 2011. [Online]. Available: http://www.rforge.net/Rserve/doc.html. [Accessed 13 9 2018].

[25] S. Urbanek, "RServe," 2010. [Online]. Available: http://www.rforge.net/Rserve/. [Accessed 05 July 2010].

[26] Generalitat de Catalunya, "DOGC núm. 5359 - 15/04/2009," 2009. [Online]. Available: http://www.gencat.cat/diari/5359/09082146.htm. [Accessed 13 9 2018].

[27] Oracle, "Oracle Weblogic Server," 2010. [Online]. Available: http://www.oracle.com/technetwork/middleware/weblogic/overview/index.html. [Accessed 11 11 2010].

[28] Apache Software Foundation, "Apache Struts," 2018. [Online]. Available: http://struts.apache.org/. [Accessed 13 9 2018].

[29] C. Cavaness, Programming Jakarta Struts,, O'Reilly Media, 2004.

[30] A. Billington, "external tables in oracle 9i," 6 2007. [Online]. Available: http://www.oracle-developer.net/display.php?id=204. [Accessed 20 10 2012].

[31] D. Moore, "SQL Loader," 2003. [Online]. Available: http://www.oracleutilities.com/OSUtil/sqlldr.html. [Accessed 22 10 2012].

[32] ORACLE-BASE.com, "XMLType Datatype In Oracle9i," 2012. [Online]. Available: http://www.oracle-base.com/articles/9i/xmltype-datatype.php. [Accessed 20 10 2012].