

Chapter 13

Scaffolding the OEEU's Data-Driven Ecosystem to Analyze the Employability of Spanish Graduates

Andrea Vázquez-Ingelmo
University of Salamanca, Spain

Juan Cruz-Benito
University of Salamanca, Spain

Francisco J. García-Peñalvo
University of Salamanca, Spain

Martín Martín-González
Technical University of Madrid, Spain

ABSTRACT

This chapter outlines the technological evolution experimented by the Observatory for University Employability and Employment's information system to become a data-driven technological ecosystem. This observatory collects data from more than 50 Spanish universities and their graduate students (bachelor's degree, master's degree) with the goal of measuring the factors that lead to students' employability and employment. The goals pursued by the observatory need a strong technological support to gather, process, and disseminate the related data. The system that supports these tasks has evolved from a standard (traditional) information system to a data-driven ecosystem, which provides remarkable benefits covering the observatory's requirements. The benefits, the foundations, and the way the data-driven ecosystem is built will be described throughout the chapter, as well as how the information obtained is exploited in order to provide insights about the employment and employability variables.

DOI: 10.4018/978-1-5225-4944-4.ch013

1. INTRODUCTION

The Observatory of University Employability and Employment (also known as *OEEU* using the Spanish initials for *Observatorio de Empleabilidad and Empleo Universitarios*) <http://www.oeeu.org/>, is an organization composed by researchers and technicians who work together from different parts of Spain with a unified methodology. The purpose of the Observatory is to produce, analyze and spread information and insights regarding the graduates' employability and employment in Spain. The Observatory is under the direction of the UNESCO Chair in University Management and Policy (based in the Universidad Politécnica de Madrid, Spain) and it relies on the mentoring of an Expert Council composed of national and international academic and university experts. This project is also developed in collaboration with the "La Caixa" Foundation, the Conference of Rectors of Spanish Universities (CRUE) and the GRIAL Research Group of the University of Salamanca (Peñalvo et al., 2012).

The Observatory's vision is to become the information reference for understanding and exploiting knowledge about variables related to employability and university employment and its behavior. To reach this vision, the Observatory has the following goals (Michavila, Martín-González, Martínez, García-Peñalvo, & Cruz-Benito, 2015; Michavila, Martínez, Martín-González, García-Peñalvo, & Cruz-Benito, 2016):

- To understand the evolution of the employability and employment, and its characteristics related to university graduates.
- To develop a system and a uniform methodology for measuring indicators about employability and employment of graduates.
- To generate information on the employability and employment of university comparable between regions, branches of study (knowledge areas) and professional profiles, among others.
- To support the development of strategies and employment policies for universities, basing it on well-founded studies and information.
- To understand the mechanisms and actions that use the Spanish universities to promote employment and employability of their graduates.
- To provide information to individual universities to adjust their academic supply and training demands to the labor market based on reliable data.

These objectives seek to resolve the lack of public information (and its analysis) regarding employment and university employability. To achieve them, the Observatory is developing, implementing and exploiting a series of data-driven products (Patil, 2012).

This data-driven approach helps the Observatory gain knowledge and wisdom from the gathered data. The data gathering procedure of the Observatory and its storage are useless if there are no other procedures to generate knowledge from raw data. Large volumes of data do not provide knowledge and wisdom by itself, but data is the base of the taxonomy of knowledge (Zeleny, 1987), and that is why it is also the base of the Observatory's system. In order to act wisely, it is necessary to have knowledge, information and data about the tasks to be solved (Zeleny, 1987).

The Observatory's studies are supported by an information system implemented to accomplish the organization's main goals. As referenced before, the Observatory products, as well as the organization itself are data-driven (Patil & Mason, 2015). This means the information system built for the Observatory

has a series of characteristics and components that facilitate the data collection, analysis, presentation and exploitation, as it will be explained below.

The first Observatory's study was released in 2015 and its goal was to collect and analyze information about degree students who graduated during the 2009-2010 course. Currently in 2017, the study is now in its second edition. The target of this second study are the university master's degree students (graduated in the 2013-2014 academic year).

However, the magnitude (a growing volume of data) and the continuity of these studies, as well as the broad vision of the Observatory, makes the traditional information systems not enough. As it will be detailed throughout this chapter, the challenges of the Observatory need the support of a more powerful system. That "powerful system" could be consider as a *data-driven* or *information-driven technological ecosystem*, because this kind of collaborative environment potentially fits the Observatory requirements.

A technological ecosystem can be seen as a set of different components connected through information flows in a physical environment that supports such flows (García-Holgado, García-Peñalvo, Hernández-García, & Llorens-Largo, 2015). Ecosystems make possible the provision of new and improved services that isolated tools or systems cannot be able to provide in order to solved knowledge and information management challenges inside any kind of institution or company (García-Peñalvo & Garcia-Holgado, 2016). Moreover, technological ecosystems provide better support to the management of information and knowledge in heterogeneous environments (García-Holgado, Cruz-Benito, & García-Peñalvo, 2015; Garcia-Peñalvo & Garcia-Holgado, 2016).

In the case of the Observatory, the implementation of a technological ecosystem should be made considering a data-driven approach, resulting into a "data-driven ecosystem". This approach implies the incorporation of the characteristics of a technological ecosystem in addition to the data-driven philosophy. The purpose of a data-driven ecosystem is to create a collaborative environment to gather, process, analyze and disseminate data, in order to make evidence-based decisions by enabling the horizontal interaction of different users, stakeholders, systems and tools. These, among others, are the characteristics that make a data-driven ecosystem suitable for the Observatory.

In addition to the above, the data-driven ecosystem provides the adequate features to support the Observatory's data-as-a-service (DaaS) approach. This means that the components within the ecosystem themselves, as well as other users, can consume data from other components on-demand and also produce and communicating data to other components (by encouraging the independence between components, their decoupling and enabling interoperability (Terzo, Ruiu, Bucci, & Xhafa, 2013). According to this philosophy and working procedure, the data-driven ecosystem stablishes a data-as-a-service approach that enables the consumption of the Observatory's raw or processed data, independently of the platform or location of the consumers and only depending on the consumer's goals and the permissions available in the ecosystem.

In 2015, the first implementation of the Observatory's information system faced a series of technical issues with keeping in mind the meeting of requirements at that moment. Later, considering the vision of becoming a reference information source, was required to make important technological changes to accomplish it. As the OEEU evolves into a fully data-driven organization, it deals with a lot of difficulties regarding the collection and sharing of heterogeneous data from different information sources, as well as it experiences the challenges associated to knowledge management and scalability. These difficulties could be addressed through a collaborative environment like an information ecosystem that could allow the management of heterogeneous data and knowledge (Cruz-Benito, García-Peñalvo, & Therón, 2014;

Scaffolding the OEEU's Data-Driven Ecosystem

García-Holgado, Cruz-Benito, et al., 2015) and could report higher levels of scalability because of the independence and low decoupling of the components involved in the ecosystem.

The purpose of this chapter is to outline the current Observatory's technological ecosystem, to describe the evolution process from a monolithic information system to a data-driven ecosystem and to present the benefits and results derived from this transition.

The content is organized following the next structure: section 2 provides a description of the different technical and organizational challenges associated to the OEEU mission and their system's goals; the third section presents the system's transition to a new version that improves previously identified issues and weakness, as well as the data structuration principles, the users involved and the technologies used; section 4 outlines the results of the Observatory and the outcomes of the designed data-driven ecosystem. Finally, the fifth section discusses the problems solved, the issues to be solved and the next challenges in the short to medium term, followed by the sixth and last section, where the conclusions of this work are presented.

2. THE PROBLEM

The Observatory of Employability and Employment faces a series of challenges and considerations to perform its work and carry out its mission.

The first challenge faced to become an information source reference are the design of data collection and data processing procedures. All the organization's purposes and products are backed by data, so data can be seen as the backbone of the Observatory. However, the possession and storage of large volumes of data are not enough to provide important outcomes. Data Analytics techniques are necessary to make these datasets valuable and to reach insights (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011) regarding the employment landscape in Spain (in this case). In brief, the Observatory needs to be able to properly support the data collection and its analysis, i.e. having efficient and automatic gathering and computation methods that enable people to make evidence-based decisions regarding employability and employment.

Nevertheless, having large volumes of data brings great responsibilities, even more if the data is considered as sensitive. The main sources of information are Spanish universities and their students; and universities, as data providers and customers of the insights provided by the Observatory, rely on the organization to keep their (and their students') data safe and anonymized. If the Observatory does not meet these conditions, universities will not trust in its projects and they will not want to be involved in them (potentially leading to the project's death by *data-inanition*).

Considering these conditions, data should be collected and stored complying the Spanish Personal Data Protection Act – generally referred to as LOPD (Ley Orgánica 15/1999, 1999). Regarding the anonymization and the Spanish law, the Observatory cannot allow the identification of individuals and universities behind the results of the study, as they will be exposed to the general audience. Anonymization and encryption techniques are needed to stick to these requirements.

In addition to the foregoing, the amount of data managed by the Observatory follows a growth trend. Nowadays, it handles more than 700 variables related more that 180000 students (these amounts are an addition of the variables and students involved in the first and second study performed by the Observatory). This is a challenge to consider, because the success of the project and the organization itself mainly remains in the capacity of handling this information in a proper way. The Observatory should

keep developing its work without became flooded by the amount of data held, and this require high levels of scalability. Scalability could be a difficult issue to manage in any organization or (eco)system, but the benefits in the medium-long term are worth it.

Finally, the Observatory needs to spread the results once completed the different studies carried out. Given the importance of the employment and employability fields for the society, it is necessary to show and exploit the knowledge produced by the organization's activities in a proper way to reach all kind of audiences and stakeholders. This knowledge must be disseminated to anyone that aims to understand the Observatory's mission as well as its studies' results. Visualization and interaction methods can help solving this challenge, as it will be shown below.

Moreover, not only the society, but another similar organizations could take advantage of the Observatory results and knowledge bank (the knowledge and extra information that support the Observatory's activities). In this case, the communication methods for this kind of outcomes are also essential. These issues will be addressed in a deeper way in the following sections.

3. TRANSITION TO A DATA-DRIVEN TECHNOLOGICAL ECOSYSTEM

3.1 The Transition Itself

The Observatory owns a significant amount of information needed to achieve the goals of analyzing and understanding the factors involved in to became more employable (employability) or in the fact of get a job (employment). Regarding that, it is necessary to build the Observatory's system focusing on the effective use of large volumes of data. Not only that, but also being capable of spread and disseminate the collected information and knowledge among the system's users as well as to other external systems, so they can interact with the information collected and processed by the Observatory to reach deeper insights about employability and employment. It is important to consider that these users will have different requirements and the Observatory needs to support all of them.

After the first implementation of the Observatory's information system in 2015, its main features and functionality were the following (Michavila et al., 2015):

- **Data Collector:** This feature allowed the data gathering from different sources. The two main information sources are described below:
 - **Universities:** The participant universities provided administrative data about their students. To accomplish the storage of the administrative data, the system allowed the universities to upload CSV files containing the information required. When uploaded, the files were validated -in a semi-automatic way- by the Observatory's staff before its definitive storage, fixing the possible errors incorporated in the initial files.
 - **Students:** The system included custom questionnaires based on the administrative data collected from the universities. The questionnaires were the method used to collect employability and employment information about the graduates involved in the study. The answers given by the students were stored in the system and completed the universities' administrative data previously described, as well as the Observatory's knowledge bank, obtaining a larger and more interesting dataset to analyze.

Scaffolding the OEEU's Data-Driven Ecosystem

- **Data Analyzer:** The main goal of this functionality was to help the analysis processes. It provided automated methods to obtain results from the gathered data. The data analyzer retrieved records from the persistent storage and processed (in a batch way) it to obtain simple statistics, derived variables and other results.
- **Interoperability:** The first implementation of the system was not prepared to support high interoperability levels between its components. On the other hand, there were neither methods to exchange information between the Observatory's system and other information systems, as it was not a priority at that time.
- **Information Dashboards:** Other main feature of the Observatory's system, supported by the data collector and the data analyzer, was the implementation of information dashboards. These dashboards allow the different users of the system to have a broad view of the results of the study through information visualizations. To make this possible, the dashboards were implemented following some basic Visual Analytics and Data Visualization principles, concepts and techniques (Keim et al., 2008; Keim, Kohlhammer, Ellis, & Mansmann, 2010). On the other hand, universities also had access to a basic dashboard with information tables to keep track of their own student's participation and other basic results.

These features satisfied the requirements at that moment (2015), giving the system the capacities to support the first edition of the Observatory's study with a simple administrative data collector and customized questionnaires for the students, as well as dashboards to show the general results of the study.

One of the principal issues were the performance of the data analysis, which made the user experience less satisfactory. Although pre-calculated statistics were used to reduce the response times in the public portal and dashboards, this solution was not flexible and, even more, it was not scalable.

As explained before, the public portal held a significant amount of visualizations representing the general results of the study, which provided general employment and employability information to anyone who accessed the results' section. However, universities were not able to have these visualizations and individual results applied to their own administrative and graduates' data. They only had the possibility of downloading the data from the system and analyze it on their own. This was an important issue and a valuable feature to be added in the future, but the performance of the data analyzer and the low levels of interoperability between the system's components made it not an easily addressable challenge.

The absence of high levels of interoperability was also a problem due to the organization's mission of analyze the data gathered and share the produced knowledge to external stakeholders. Increasing these interoperability levels could allow other information systems to take advantage of the Observatory's data, creating a collaborative knowledge environment and reporting more benefits to the project and society.

After analyzing all the issues found behind the first implementation of the Observatory's system, it was clear that the system must be upgraded to prepare the environment for future studies and functionalities. This was when the transition to a data-driven ecosystem began, considering the similarities between the Observatory's requirements and the problems addressed by this approach.

It is important to consider the challenges and priorities behind the construction of a digital ecosystem focused in knowledge discovery through data analysis and exploitation in order to obtain valuable insights from the domain data (Chajri & Fakir, 2014; Touya & Fakir, 2014). These priorities especially involve the interoperability between the series of components within the ecosystem, as well as the flexibility to be used by different stakeholders, and the scalability regarding to data management and data analysis (Garcia-Peñalvo & Garcia-Holgado, 2016). As referenced before, if high levels of interoperability are

achieved, the components could be able to evolve individually and collectively, and they even could add or change their own functionalities without affecting the rest of the components. Moreover, additional components, both internal or external, could be connected to build up a broader technological ecosystem without major difficulties in a transparent way (García-Holgado & García-Peñalvo, 2014a, 2014b).

Those characteristics met the Observatory's philosophy; employment and university employability fields are constantly evolving, and the Observatory needs to adapt their studies to get the maximum amount of information and analyze it in the best possible way. The Observatory needs to collect, process, analyze and present data, and in order to support those activities, as presented above, a data collector, data analyzer and data dashboards functionalities have been implemented through a series of software components. These components consume information from the persistent storage of the Observatory and also consume data generated by each one of them. It is clear that every component should be independent and need to have well-defined tasks, but also they must collaborate to achieve the goals of the whole system. Information flows between the components are important in order to reach the collaboration of all them, and promoting high interoperability levels is a common solution to decouple the components while they continue collaborating in a transparent way.

For all these reasons, it has been concluded that the data-driven ecosystem approach would report very important benefits to the Observatory system purposes. By implementing a data-driven ecosystem, the Observatory's technological environment can evolve following new requirements and providing new functionalities or services by the addition or modification of components, thanks to the high cohesion of the components but also to their loose coupling.

Taking all these points into account, the system has been upgraded in order to begin a transition to a data-driven ecosystem and accomplish the Observatory's new requirements due next studies' editions, as well as to fix other issues identified after the 2015 study and with the authors' previous work in software architectures that gather and manage data in different contexts (Cruz-Benito, Borrás-Gené, García-Peñalvo, Fidalgo Blanco, & Therón, 2015; Cruz-Benito, García-Peñalvo, et al., 2014; J. Cruz-Benito et al., 2016; Cruz-Benito, Therón, et al., 2014; Cruz-Benito, Therón, & García-Peñalvo, 2016; Cruz-Benito, Therón, García-Peñalvo, & Pizarro Lucas, 2015; García-Peñalvo, Cruz-Benito, Maderuelo, Pérez-Blanco, & Martín-Suárez, 2014; García-Sánchez, Cruz-Benito, Therón, & Gómez-Isla, 2015).

This upgrade took place during the development and beginning of the second study conducted by the Observatory (2016-2017). Among the different feature additions and modifications, the following could be highlighted:

- **Data Collector:** The component regarding the data collection has been upgraded to solve the issues encountered after the first implementation of the system. Despite the technical changes, the information sources have not varied:
 - **Universities:** As in the 2015 edition, the data gathering starts with universities sending its administrative information. One of the issues of that first edition was the need of manual validations by the Observatory's staff before considering the data as correct. This manual validation procedure slowed down the gathering process, since it was not automatized and did not give guarantees of the proper correction of all the errors found. For that reason, the administrative data collector has been upgraded to include a fully-automated data validation stage before the information is persistently stored into the system. The validation stage avoids manual checking procedures by analyzing the CSV files just after the universities have uploaded it, and presenting them a detailed report containing the exceptions and its

Scaffolding the OEEU's Data-Driven Ecosystem

explanations. Universities are responsible of fixing the errors found and upload their new and corrected files afterwards. Once the system receives a file without any errors, the data passes another stage of cleaning and structuring to keep the Observatory's bank of information organized, as it will be briefly explained in the next subsection.

- **Students:** The student's data collector has also experienced an upgrade. Rather than fix issues from the first edition, some changes had been added in order to make the current component more valuable. The data is collected (again) through questionnaires, but one of the main upgrades includes the improvement of the questionnaires themselves by creating different versions that will be displayed depending on the student that fills the questionnaire (some kind of adaptivity), in pursuance of an increase of the completion rate. Another upgrade involves the collection of variables generated from the student's interaction with the questionnaires (technically referred to as *paradata* (Stieger & Reips, 2010). These adaptive questionnaires themselves could lead to future research about this topic (research using A/B methodologies, etc.) (Cruz-Benito et al., 2017). As outlined in the previous bullet, the students' answers are also structured and cleaned before saving it permanently.
- **Data Analyzer:** The data analyzer has been modified to increase its performance and support more complex operations. The data retrieval process is made through the same methods as in the old version, against the Observatory's database, excepting the data calculations are performed on-demand (and not in a batch way) due to the increased efficiency through an in-memory analysis approach. Once the data is retrieved, it is loaded into memory, where the processing and analysis take place. In-memory analysis allows the datasets to be processed more efficiently, reporting significant increases of performance (Zhang, Chen, Ooi, Tan, & Zhang, 2015). This enhancement opens the capacities of the whole system, as most components could consume information generated by the data analyzer.
- **Interoperability:** A data-driven ecosystem should be able to exchange and exploit data along the components that form part of it. Interoperability reports several benefits to the Observatory's system and improve scalability, communication and collaboration between its elements. To achieve these goals, an interoperability component has been introduced to the system. This component was implemented by the incorporation of a REST (Fielding & Taylor, 2002) API (Application Programming Interface). Different data flows can be found within (and against) the system, so the API is designed to contemplate and give support to all of them. The design of this API brings to the Observatory's database isolation properties, as transactions won't interfere with each other considering the secure behavior of the API requests. This component allows the data to be available for users and other components through network, making a data-as-a-service approach possible (Terzo et al., 2013). It also allows the creation of methods for the users to download their entire datasets in different open format files (like CSV, etc.) on-demand.
- **Information Dashboards:** In general, the system's information dashboards and visualizations are the main beneficiaries of the transition to a data-driven ecosystem. The poor performance level of the older version of the system led to difficulties regarding the provision of individual dashboards to each participant university. The enhancement of the communication between components and the improvement of the data analyzer made possible the sending of real-time data to the information dashboards (as the data is retrieved and processed on demand). In addition, the good performance of this procedure makes the user experience more satisfying.

The backbone of the transition from a monolithic system to a data-driven ecosystem is the interoperability component, not only because it can be seen as a nexus between several information (eco)systems, but also because it made possible the scalability of the information dashboards and other components of the system. This interoperability component enables the information dashboards to consume processed data from the data analyzer component, decoupling them from other system's components and supporting the data-as-a-service approach. Having high decoupling levels facilitates the evolution of every component individually, without interfering with each other, but also collectively, meeting the characteristics of a technological ecosystem (García-Holgado & García-Peñalvo, 2014a, 2014b; García-Holgado & García-Peñalvo, 2014a, 2014b).

All the changes and functionalities introduced in this transition are resumed in *Table 1*, where the 2015 information system components and characteristics are compared with the 2017 data-driven technological ecosystem.

Finally, Figure 1 shows the conceptual view of the Observatory's data-driven ecosystem.

To complete the transition overview, the next subsections briefly describe the data structuration referenced before, the system's users, and the different technologies involved in the new ecosystem.

3.2 Data Structuration

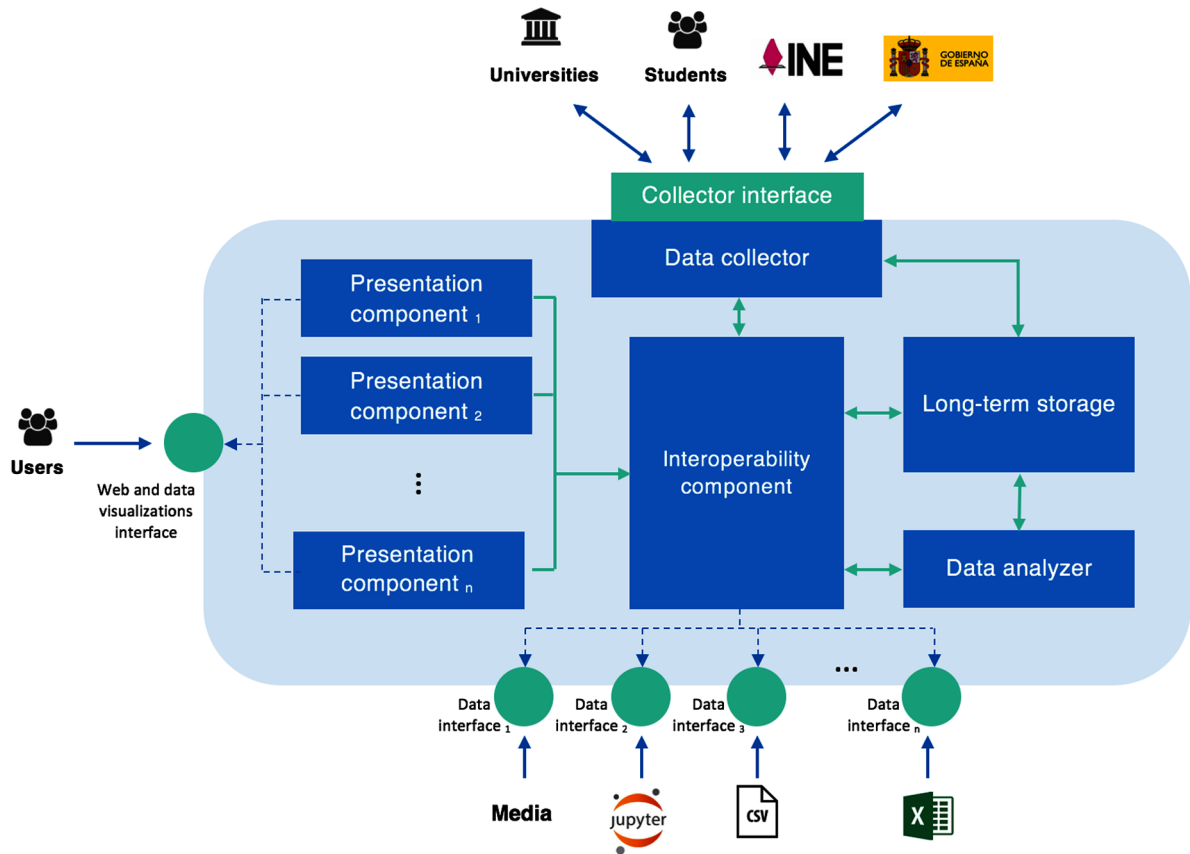
Considering that the data is the driver element of the system, it is required to organize and structure it to perform its processing and analysis properly, reporting benefits for all the Observatory consumers. This structuration procedure is transparent for the users and it takes place at the system's backend.

Table 1. Summary of changes in the transition from a standard information system to a data-driven technological ecosystem

	OEEU's 2015 Information System	OEEU's 2017 Data-Driven Ecosystem
Data collector	<ol style="list-style-type: none"> 1. Procedures to collect administrative data from universities. 2. Subsystem to generate questionnaires and collect data from the students' answers. 	<ol style="list-style-type: none"> 1. New stages in the gathering procedures: on-demand data validation and data cleaning. 2. Subsystem to generate questionnaires and different versions of them that could be displayed depending on the student profile that participates (supporting A/B testing). 3. Collection of <i>paradata</i> from the students involved in the study.
Data analyzer	<ol style="list-style-type: none"> 1. Low levels of performance 2. Simple analysis 	<ol style="list-style-type: none"> 1. Increase of performance due to the in-memory data computation 2. Support of more complex analysis
Interoperability	Not contemplated in requirements.	Addition of an interoperability component to communicate the different components within the ecosystem and external systems or users through data flows, following the data-as-a-service approach.
Information dashboards	<ol style="list-style-type: none"> 1. Visualizations for general results in the public website and personalized data tables for each universities' results in the system intranet. 2. Pre-calculated results. 3. Coupling between presentation and analysis components. 	<ol style="list-style-type: none"> 1. Adaptation of the visualizations and general dashboards to show each universities' results in their private dashboards. 2. Results calculated on-demand based on the request. 3. Higher levels of decoupling due to the interoperability component developed.

Scaffolding the OEEU's Data-Driven Ecosystem

Figure 1. OEEU's ecosystem's components and relations



The principles followed to accomplish the data organization and structuration are technically named *tidy data* principles (Wickham, 2014). The Observatory's bank of information is organized in such a manner that, as Wickham points out, each variable forms a column, each observation forms a row and each type of observational unit forms a table. The raw data gathered from the different information sources is cleaned before its storage into the system's persistence layer sticking to these rules.

The principal benefit is the avoidance of messy datasets and the need of additional computations each time information is requested. For that last reason, tidy datasets also report performance gains as these unnecessary pre-processing operations are omitted.

3.3 System Users

There are a series of users of the Observatory's system, categorized by their needs. These users can be common users of the Observatory's public website (general audience), the graduates involved in the Observatory's studies, the Spanish universities that participate in the data gathering and consuming the knowledge and information generated, the data analysts (both external or internal to the organization) and other information systems that take advantage of the Observatory's components and information.

3.4 Technologies Used

3.4.1 Core System

The system has been implemented on Django, a Python web framework (Django Software Foundation, 2015; Holovaty & Kaplan-Moss, 2009). The election of Django has been backed up by the simplicity, flexibility and the community behind this framework.

On the other hand, the API has been built with the toolkit Django REST Framework (<http://www.django-rest-framework.org/>) and one of its variants: Django REST Pandas (<https://github.com/wq/django-rest-pandas>), which allows the API to serve Pandas dataframes (McKinney, 2011) to the client side. These frameworks meet the requirements of the project as they keep simplicity and good performance levels and provide a series of authorization and authentication methods for limiting and securing the access to the API, which is composed by different layers to keep the data model, the security and the logic of the system separate from each other.

3.4.2 Data Persistence

The long persistence technology used is MariaDB (<https://mariadb.com/>), a database management system derived from MySQL. With this technology, the main goals to achieve are to keep all of the Observatory's data safe, structured and well-organized, as well as obtain satisfactory response times on query processes.

3.4.3 Data Processing and Analysis

As already outlined above, the data analysis and processing take place at the backend of the system. These backend computations are made through Python Pandas (McKinney, 2011) because it allows in memory column analysis, reporting huge performance benefits.

All the computed data served by the Observatory's new engine is processed before sending it to the user who requested it, avoiding additional computational load on the user's side. Using this method, the system's information dashboards can show real-time data, unlike in the previous version of the system, where the statistics were pre-calculated, leading to flexibility and scalability problems.

3.4.4 Data Visualizations

The information dashboards have been built using D3, a Javascript library for visualizing data (Bostock, Ogievetsky, & Heer, 2011). D3 provide the appropriate methods to visualize the Observatory's bank of information, accomplishing the requirements and necessities of the users.

3.4.5 Data Exploitation

The interoperability component makes easy the external data retrieval and data exploitation through different tools independent of the system.

For example, data analysts can retrieve information from the Observatory's API and process the raw data using external tools like Jupyter notebooks (Kluyver et al., 2016), which could be very useful for publishing results, considering their capacity to show live code and outcomes even without being ex-

Scaffolding the OEEU's Data-Driven Ecosystem

ecuted and the reproducibility feature it provides regarding to data analysis processes. These notebooks take advantage of the Observatory's API, facilitating the data retrieval process and hiding details of the system's persistent storage.

The ecosystem also has methods to allow universities to retrieve their students' data in CSV format, in order to use the downloaded files in other external tools like IBM SPSS, Microsoft Excel, etc.

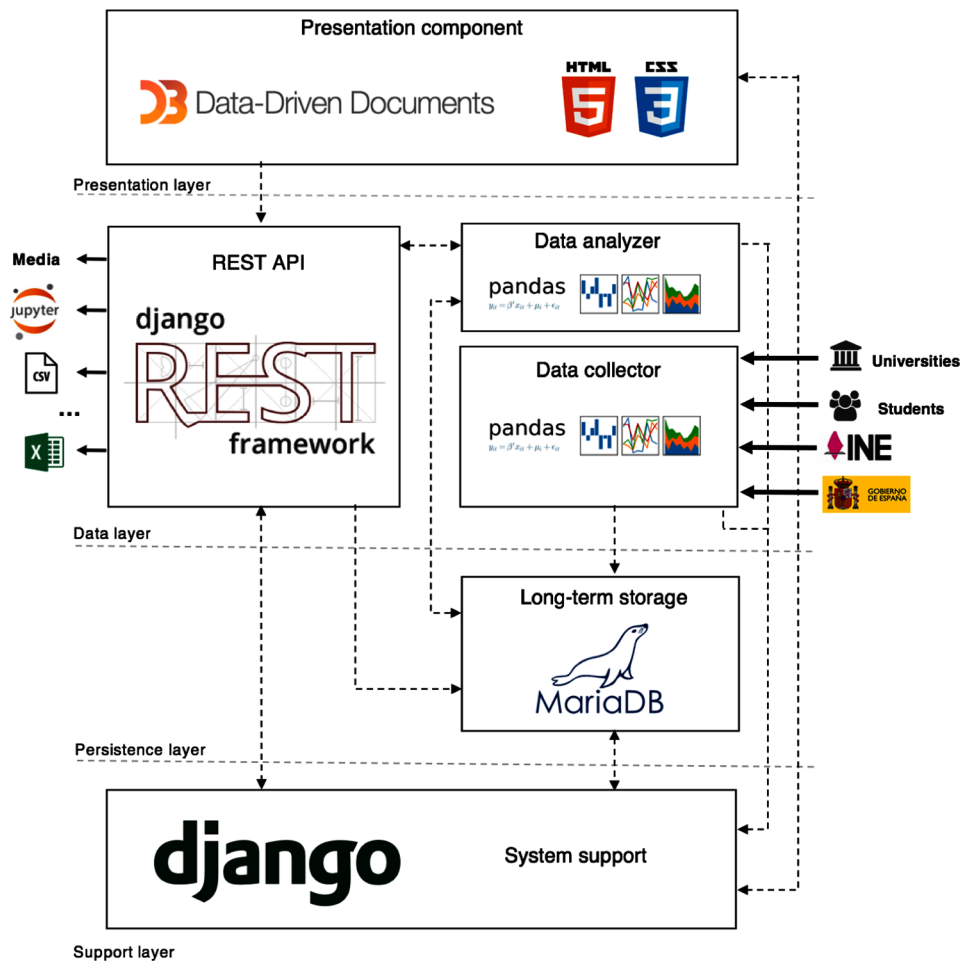
The technologies involved in the Observatory's ecosystem are summarized in Figure 2.

4. RESULTS

4.1 Observatory's Results

The Observatory for University Employability and Employment has been running since 2013, and started to collect degree student's information between the end of 2014 and the beginning of 2015, merging both

Figure 2. Summary of components and technologies that compose the Observatory's data-driven ecosystem



universities' administrative data and students' information collected through questionnaires. This data gathering corresponds to students that graduated in bachelor level in the 2009-2010 course.

At the end of 2016, a new period for data collection started, beginning with universities sending administrative information about students who finished their master's degree in the 2013-2014 academic year. As well as in the 2015 study, another data gathering process through questionnaires began in the early months of 2017, complementing the administrative information from universities.

The following results were reached during the Observatory's 2015 study (Michavila et al., 2016):

- The Observatory involved 49 Spanish universities, both public and private ones.
- The administrative data provided from all of these 49 universities reached at the end of the data collection stage 134,129 records about graduated students in the 2009-2010 season. Information about gender, sex, nationality, parents' studies, place where students live, average score in studies, mobility, internships, etc. are included among the total of data variables given (35 altogether).
- Moreover, of these 134,129 students, 13,006 started to answer their questionnaires, which is the 9,70% of the total registered students expressed as a percentage. Finally, 9,617 of them finished it (7,17% of the students in this case), giving a finalization rate of the questionnaires of 73,94%. The information provided by the completed questionnaires include around 400 variables regarding studies and employment. These variables keep information about the type of employment, relation with their jobs and their studies, wage, competencies, other studies done, opinions about the university and teaching methods, ratings about their studies, languages spoken, etc.
- Also, the Observatory keeps a knowledge bank with information related to Spanish degrees, information about branches of knowledge, as well as countries' data (economic, demographic, etc.), fields of employment and other relevant information about companies or employers. This information has been organized by the Observatory and can be useful to reach wider results data analysis.

On the other hand, the Observatory's 2017 study has reached other promising results during these last months of data gathering:

- In this case, The Observatory has involved 51 Spanish universities so far, also public and private. The difference on the number of participants between the two studies has its explanation in the fact that some of the universities that collaborated in the first study had decided not to participate in this last one and others that did not participate were enrolled in this new study. In the new study, universities added administrative data about its own master's graduated students.
- This time, the administrative data provided from these 51 universities involved 47,822 records about master students.
- Of these 47,822 students, approximately 6700 started questionnaires and 5200 finished them, making the questionnaire's finalization rate reach the more than 77%. The completion of the master's degree (MsC) questionnaires provides more than 200 variables related, again, ranging from information about their jobs, studies, competencies and other employment and university studies attributes.
- As in the 2015 study, variables given by the Observatory knowledge bank are used to allow deeper data analysis.

4.2 Results Regarding the System's Evolution

Regarding the results of the Observatory's information ecosystem itself, including the transition to a more powerful architecture supported by the concept of digital ecosystems and data-driven culture, the technological achievements reached are remarkable:

- The Observatory provides tools to help the universities to send its administrative information, as well as validation methods that give real-time feedback about possible errors within uploaded files.
- The Observatory has a system to generate personalized questionnaires for the involved students. The personalization is based on the student responses and profile. The personalization goes from the questionnaires' style to its interaction methods, among other changes. It also has procedures to gather information about the students' interaction with the questionnaires that could lead to future researches about usability and adaptivity.
- The Observatory has implemented components to analyze the collected data on-demand, achieving low response times even when computing large datasets.
- Finally, regarding the data presentation, the Observatory has a system for data visualization that handles the information for different stakeholders (universities, students or staff from the Observatory itself, among others). These presentations vary depending on the stakeholder.
 - The staff from the Observatory can access to all the data collected in order to obtain general results, as well as to manage individual universities' results or to monitor the questionnaires participation data.
 - Universities have access to personalized dashboards with information tables and visualizations adapted to its own data. This allow universities to consult effortlessly their results based on its administrative information and its students' answers.
 - The Observatory also provides public websites (<http://datos.oeeu.org>) that show the different studies overall results in an open way for any user. The websites make possible to disseminate all the knowledge discovered by the system.

The Figure 3 and Figure 4 show examples of the main Observatory's dashboards and its data visualizations, respectively. As explained before, these dashboards are also available for each participant University to check their own results.

5. DISCUSSION

Due to the upgrade of the Observatory's system, several issues associated to the complexity of the goals and technical challenges have been solved.

The Observatory's system has experienced an important performance increase in server's response, data analysis and data presentation, as this last one feature consumes data from the data analyzer. Since the nature of the project and the Observatory itself implies a continuous growth of the data amount, this was a priority issue to be resolved to avoid future problems regarding response and computation times.

Figure 3. Example of the administrative dashboard for the Observatory's staff users (contents in Spanish)

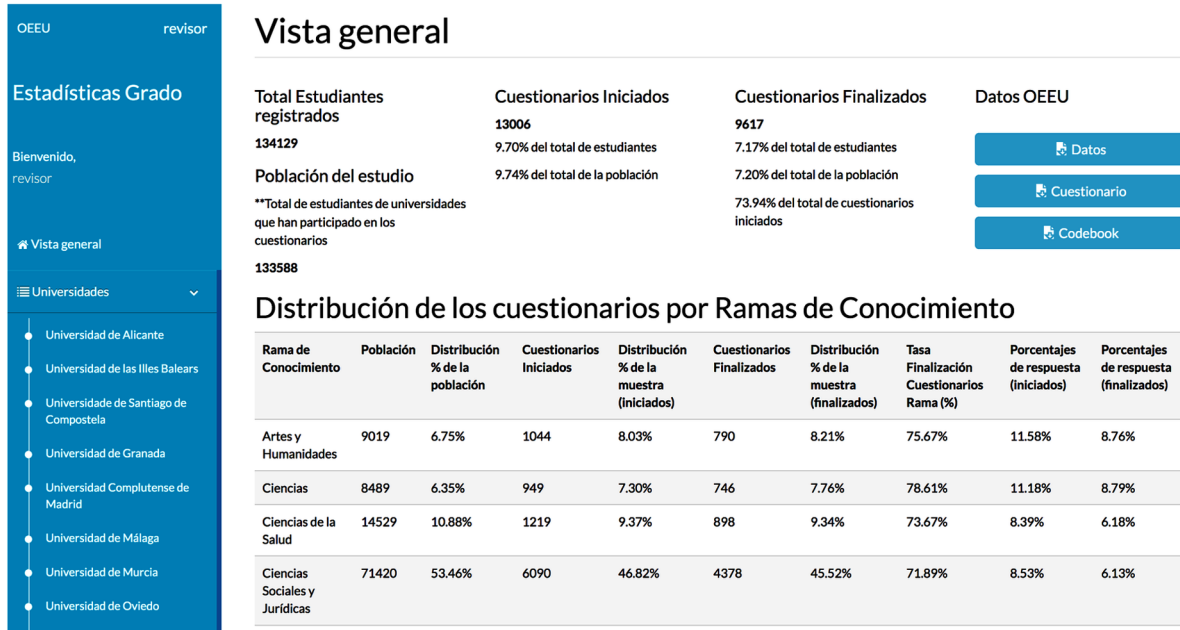


Figure 4. Example of information dashboards for the Observatory's data (contents in Spanish)



Scaffolding the OEEU's Data-Driven Ecosystem

On the other hand, the data validation stage in the administrative data collector has provided great results, accelerating the first phase of the study and giving real-time feedback to the universities about the validity of their data. Before introducing an automatic data validator, manual fixes were essential and critical, because erroneous data could introduce noise in the studies. This manual validation consumed notable human and time resources from the Observatory, and it even could not be enough to fix all errors. Having an automatic validation stage, the errors are showed directly to the universities, making them responsible of the fixes and unburdening the Observatory from this task.

However, the main results come from the increase of interoperability of the system and the transition to a data-driven ecosystem. As it was introduced at the beginning of this chapter, the Observatory deals with high amounts of data coming from different sources, being the data the backbone of the system and the basis to make evidence-based decisions. The components within the system need to communicate with each other in a proper way to exchange the results of their tasks, but they also need to maintain high levels of independence, following technological ecosystems' guidelines and creating a data-as-a-service environment.

The data-driven ecosystem implemented allow software components to evolve individually and collectively, that made possible the addition of new functionalities and features without major difficulties. It reports remarkable benefits regarding communication between components and it reduces integration issues. Other of the results given by the interoperability and independence of the components is the enhancement of the information dashboards and the communication with third-party tools and systems. Data flows between visualizations and other components of the ecosystem enable dashboards to show more complex information, and even filter it depending on the user.

Although all the problems encountered after the first edition of the study have been covered, the needs of the Observatory are not static, and they are evolving continuously. The data-driven ecosystem implemented is a good basis to begin addressing more ambitious challenges. The decoupling of the components, especially the decoupling between the data analyzer and the dashboards opens many doors regarding information visualizations. Regarding that, some next challenges could include the evolution of the API responses to a graph approach for complex data (Vázquez-Ingelmo, Cruz-Benito, & García-Peñalvo, 2017), like it is being developed by main organizations as Facebook (i.e. with the GraphQL project <https://facebook.github.io/graphql/>).

One of the next challenges for the medium- or long-term is to develop linked data views for the visualization on the information dashboards. The philosophy behind linked visualizations lies in the creation of several simpler views (instead of creating one complex view) and their linkage, so that when the user interacts with one view, the other views will update and show the results of such an interaction (Wills, 2008). The Observatory owns a lot of information collected in each study, and it is difficult to reach valuable insights without the support reached through visualizations and interaction. Linked data views would make easier for the users to understand the Observatory's results. Also, could not be discarded the addition of complex data visualizations for the most complex problems regarding the information kept by the Observatory (multidimensional problems, etc.).

The data-driven ecosystem approach also allows the dissemination of information and knowledge gained by the Observatory among other external systems. Media, journals, data analysts etc., could link their own systems or tools to the Observatory's ecosystem to obtain results and insights of employment and university employability.

6. CONCLUSION

The Spanish Observatory for University Employability and Employment is generating, analyzing and disseminating information about the employability and employment of university graduates in Spain through a unified methodology. To support the data collection, as well as its analysis and dissemination, an information system was built in 2015 to accomplish the Observatory's goals and technological requirements. However, giving the challenges of the Observatory regarding its mission and information requirements, the traditional information system built has been transformed into a data-driven ecosystem. As it was presented throughout this chapter, this data-driven technological ecosystem presents a collaborative environment between software components and users that supports adequately the Observatory's vision and mission.

Technological ecosystems empower this kind of collaborative environments, enabling the connection of different components through data flows, resulting in a series of individual components with well-defined task working together to achieve the ecosystem's goals. They also allow the modification and update of components without affecting the rest of the ecosystem's elements, and makes easier the creation of new data flows, both internal and external to the ecosystem. On the other hand, data-driven philosophy focuses the attention on data in order to support decision making.

The combination of these two concepts make a data-driven ecosystem a powerful and suitable approach for the Observatory. This combination matches well the data-as-a-service approach implemented in the project, due to the interoperability, independence and decoupling gained through the transition to a data-driven ecosystem, allowing data to be served on-demand to different components and users.

With its data-driven ecosystem, the Observatory is now able to discover and understand better the variables that influence the employment and employability in graduate students. It also promotes the creation of data flows to disseminate the knowledge and wisdom gained during the analysis to internal and external users, different stakeholders and third-party information systems or tools, as well as it supports the evolution of its own components to meet new requirements.

ACKNOWLEDGMENT

The research leading to these results has received funding from “la Caixa” Foundation. Also, the author Juan Cruz-Benito would like to thank the European Social Fund and the *Consejería de Educación* of the *Junta de Castilla y León* (Spain) for funding his predoctoral fellow contract. This work has been partially funded by the Spanish Government Ministry of Economy and Competitiveness throughout the DEFINES project (Ref. TIN2016-80172-R).

REFERENCES

- Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2301–2309. doi:10.1109/TVCG.2011.185 PMID:22034350
- Chajri, M., & Fakir, M. (2014). Application of Data Mining in e-Commerce. *Journal of Information Technology Research*, 7(4), 79–91. doi:10.4018/jitr.2014100106

Scaffolding the OEEU's Data-Driven Ecosystem

Cruz-Benito, J., Borrás-Gené, O., García-Peñalvo, F. J., Fidalgo Blanco, Á., & Therón, R. (2015). Extending MOOC ecosystems using web services and software architectures *Proceedings of the XVI International Conference on Human Computer Interaction* (pp. 1-7). Vilanova i la Geltrú, Spain: ACM. doi:10.1145/2829875.2829923

Cruz-Benito, J., García-Peñalvo, F. J., & Therón, R. (2014). *Defining generic data collectors for Learning Analytics: Facing up the heterogeneous data from heterogeneous environments*. Paper presented at the International Conference on Advanced Learning Technologies (ICALT) 2014, Athens, Greece. doi:10.1109/ICALT.2014.108

Cruz-Benito, J., Maderuelo, C., García-Peñalvo, F. J., Therón, R., Pérez-Blanco, J. S., Gómez, H. Z., & Martín-Suárez, A. (2016). Usalpharma: A Software Architecture to Support Learning in Virtual Worlds. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, 11(3), 194–204. doi:10.1109/RITA.2016.2589719

Cruz-Benito, J., Therón, R., & García-Peñalvo, F. J. (2016). Software Architectures Supporting Human-Computer Interaction Analysis: A Literature Review. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies: Third International Conference, LCT2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings* (pp. 125-136). Cham: Springer International Publishing. doi:10.1007/978-3-319-39483-1_12

Cruz-Benito, J., Therón, R., García Peñalvo, F. J., Maderuelo, C., Pérez-Blanco, J. S., Zazo, H., & Martín-Suarez, A. (2014). Monitoring and feedback of Learning Processes in Virtual Worlds through analytics architectures: A real case. In Á. Rocha, D. Fonseca, E. Redondo, L. P. Reis, & M. P. Cota (Eds.), *Sistemas y Tecnologías de Información. Actas de la 9ª Conferencia Ibérica de Sistemas y Tecnologías de Información* (Vol. I Artículos, pp. 1126-1131). Barcelona, España: AISTI (Asociación Ibérica de Sistemas y Tecnologías de Información). doi:10.1109/CISTI.2014.6877097

Cruz-Benito, J., Therón, R., García-Peñalvo, F. J., & Pizarro Lucas, E. (2015). Discovering usage behaviors and engagement in an Educational Virtual World. *Computers in Human Behavior*, 47(0), 18–25. doi:10.1016/j.chb.2014.11.028

Cruz-Benito, J., Therón, R., García-Peñalvo, F. J., Sánchez-Prieto, J. C., Vázquez-Ingelmo, A., Martín-González, M., & Martínez, J. M. (2017). Improving Success/Completion Ratio in Large Surveys: A Proposal Based on Usability and Engagement. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies. Technology in Education: 4th International Conference, LCT 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part II* (pp. 352-370). Cham: Springer International Publishing. doi:10.1007/978-3-319-58515-4_28

Django Software Foundation. (2015). *Django Web Framework*. Retrieved from <https://www.djangoproject.com/>

Fielding, R. T., & Taylor, R. N. (2002). Principled design of the modern Web architecture. *ACM Transactions on Internet Technology*, 2(2), 115–150. doi:10.1145/514183.514185

García-Holgado, A., Cruz-Benito, J., & García-Peñalvo, F. J. (2015). *Analysis of Knowledge Management Experiences in Spanish Public Administration*. Paper presented at the Third International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'15), Porto, Portugal. doi:10.1145/2808580.2808609

García-Holgado, A., & García-Peñalvo, F. J. (2014a). *Architectural pattern for the definition of eLearning ecosystems based on Open Source developments*. Paper presented at the Computers in Education (SIIE), 2014 International Symposium on. doi:10.1109/SIIE.2014.7017711

García-Holgado, A., & García-Peñalvo, F. J. (2014b). *Patrón arquitectónico para la definición de ecosistemas de eLearning basados en desarrollos open source*. Academic Press.

García-Holgado, A., García-Peñalvo, F. J., Hernández-García, Á., & Llorens-Largo, F. (2015). *Analysis and improvement of knowledge management processes in organizations using the Business Process Model Notation*. Paper presented at the Annual Conference of the Global Innovation and Knowledge Academy. doi:10.1007/978-3-319-22204-2_9

García-Peñalvo, F. J., Cruz-Benito, J., Maderuelo, C., Pérez-Blanco, J. S., & Martín-Suárez, A. (2014). Usalpharma: A Cloud-Based Architecture to Support Quality Assurance Training Processes in Health Area Using Virtual Worlds. *The Scientific World Journal*, 10. doi:10.1155/2014/659364 PMID:24778593

García-Peñalvo, F. J., & García-Holgado, A. (2016). *Open Source Solutions for Knowledge Management and Technological Ecosystems*. IGI Global.

García-Sánchez, F., Cruz-Benito, J., Therón, R., & Gómez-Isla, J. (2015). *Designing and building systems and tools to analyze visual communications on social networks*. Paper presented at the Third International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'15), Porto, Portugal. doi:10.1145/2808580.2808629

Holovaty, A., & Kaplan-Moss, J. (2009). *The definitive guide to Django: Web development done right*. Apress. doi:10.1007/978-1-4302-1937-8

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. *Information Visualization*, 154–175.

Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering the information age solving problems with visual analytics*. Eurographics Association.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., . . . Corlay, S. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 52(2), 21.

Ley Orgánica 15/1999. (1999). Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. *Boletín Oficial Del Estado*, 298(2), 43088-43099.

Scaffolding the OEEU's Data-Driven Ecosystem

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 1-9.

Michavila, F., Martín-González, M., Martínez, J. M., García-Peñalvo, F. J., & Cruz-Benito, J. (2015). Analyzing the employability and employment factors of graduate students in Spain: The OEEU Information System. *Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality*. doi:10.1145/2808580.2808622

Michavila, F., Martínez, J. M., Martín-González, M., García-Peñalvo, F. J., & Cruz-Benito, J. (2016). *Barómetro de Empleabilidad y Empleo de los Universitarios en España, 2015 (Primer informe de resultados)*. Academic Press.

Patil, D. (2012). *Data Jujitsu*. O'Reilly Media, Inc.

Patil, D., & Mason, H. (2015). *Data Driven*. O'Reilly Media, Inc.

Peñalvo, F. J. G., Conde, M. J. R., Pardo, A. M. S., González, M. Á. C., Zangrando, V., & Holgado, A. G. (2012). GRIAL (GRupo de investigación en InterAcción y eLearning), USAL. *IE Comunicaciones: Revista Iberoamericana de Informática Educativa*, (15), 85-94.

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, 26(6), 1488–1495. doi:10.1016/j.chb.2010.05.013

Terzo, O., Ruiu, P., Bucci, E., & Xhafa, F. (2013). *Data as a service (DaaS) for sharing and processing of large data collections in the cloud*. Paper presented at the Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on.

Touya, K., & Fakir, M. (2014). Mining Students' Learning Behavior in Moodle System. *Journal of Information Technology Research*, 7(4), 12–26. doi:10.4018/jitr.2014100102

Vázquez-Ingelmo, A., Cruz-Benito, J., & García-Peñalvo, F. J. (2017). *Improving the OEEU's data-driven technological ecosystem's interoperability with GraphQL*. Paper presented at the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality, Cádiz, Spain.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1–23. doi:10.18637/jss.v059.i10 PMID:26917999

Wills, G. (2008). Linked data views. *Handbook of data visualization*, 217-241.

Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. *Human Systems Management*, 7(1), 59–70.

Zhang, H., Chen, G., Ooi, B. C., Tan, K.-L., & Zhang, M. (2015). In-memory big data management and processing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 27(7), 1920–1948. doi:10.1109/TKDE.2015.2427795