International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

Integration and Open Access System Based on Semantic Technologies:
A Use Case Applied to University Research Facet

Ana María Fermoso García, Pontificia University of Salamanca, Spain*
Maria Isabel Manzano García, Pontificia University of Salamanca, Spain
Roberto Berjón Gallinas, Pontificia University of Salamanca, Spain
Montserrat Mateos Sánchez, Pontificia University of Salamanca, Spain
María Encarnación Beato Gutiérrez, Pontificia University of Salamanca, Spain

ABSTRACT

The aim of this work is the development of an information system that, by integrating data from different sources and applying semantic technologies, makes it possible to publish and share with society the scientific production generated in the university environment, promoting its dissemination and thus contributing to the knowledge society, among others. In practice, this is the implementation of a CRIS (current research information system). This CRIS presents advanced features. On one hand it applies semantic technologies, providing a query service through a SPARQL Point, besides the reuse of shared data by exporting them in different formats. In this sense, it is also based on a European ontology or semantic standard such as CERIF, which facilitates its portability. On the other hand, CRIS also presents an alternative to the lack of a single data system by allowing data from different sources to be integrated and managed.

KEYWORDS:

CRIS (Current Research Information System), RIM (Research Information Management), CERIF (Current European Research Information Format), SPARQL Point, Ontology, Open Data, Information integration

INTRODUCTION

The University is entrusted with two main tasks, teaching and research. In fact, the fundamental nucleus of its staff is defined as TRS or Teaching and Research Staff.

As non-profit organizations, universities play a key role in society. Specifically, and in relation to their research facet, the field to which this paper refers, they have almost the duty to be accountable to society by sharing with it the results of their research and scientific production. Its aim is to contribute to social and scientific progress and, in short, to the knowledge society. Within the university, university libraries should be considered as fundamental agents involved in the protection and dissemination of the research carried out in the institution. In fact, this work has been developed in the academic environment by a multidisciplinary team made up of staff from the library and technological fields within the University.

DOI: 10.4018/IJSWIS.309422 *Corresponding Author

A CRIS (Current Research Information System), also called RIM (Research Information Management) if it is used the American denomination, is as its own acronym denotes, an information system that serves to collect and disseminate all the information related to the research activities of an institution (Abadal, 2019), the university in this case. Based on this definition and the purpose of the project, it is concluded that what has been designed is a CRIS system. However, the project goes further in its features in relation to a traditional CRIS, so it can be considered as a CRIS with advanced features, among others thanks to its use of semantic technologies.

In brief, the main objective of this work is the development of an information system that by integrating data from different sources and applying semantic technologies, allows to publish and share with society the scientific production generated in the university environment promoting its dissemination.

Background

CRIS is undoubtedly the best tool to help disseminate the university's research work. However, only a few universities have this type of system, although they should become a regular tool in the university environment and not only for their own interest and benefit, but also because they can become a system for transferring knowledge to society and justifying their activities to it (Bryant, et al., 2018). A CRIS system can facilitate the exchange of information, contact with colleagues working on related topics, internal support for decision making (Jeffery, 2006), or even for the comparison between universities and measurement of their research capacity and for the justification of their activity to the different administrations and society, to which the university must contribute with its knowledge. Therefore, there are many benefits (Ebert, 2014), but also for this reason, the quality and performance of the system implemented and of the agents involved in leading its management and development, are fundamental.

CRIS at University Context

Analyzing a couple of studies on the situation of CRISs in the Spanish and European university environment, it is seen that CRIS is also turning out to be case studies, as shown in two recent papers on the state of CRIS.

One of them, carried out by the CRUE (Conference of Rectors of Spanish Universities), analyzes the situation of CRISs in the Spanish university system (Malo de Molina, 2018) through a recent survey of some Spanish universities carried out between June and July 2018 within REBIUN (Spanish University Library Network). This report tries to see the involvement of university libraries in the work of the CRIS, but also, to analyze the areas that manage CRIS, who is in charge of their management, where they are hosted and what applications they use, how they interact with external systems, ...

The other report analyzes the management of CRIS in the international context (Abadal, 2019). This report highlights among its conclusions, that CRIS should be closely related to institutional repositories and therefore these must become true single data systems. The report also highlights that the management of a CRIS affects different agents in the university, among them and fundamentally, the university libraries in fulfillment of their research support functions.

In this work it has been taken into account the aspects included in both reports. On the one hand, with regard to the functionalities expected in a CRIS, on the other hand, from the interdisciplinary collaboration itself when developing the project, between the faculty of computer science as technological side, and the library due to experience in supporting research and its dissemination in university and sociocultural environment.

There are also other use cases related to facilitate interoperability between different university CRIS. In (Anna Clements, 2017) it is highlighted how this interoperability should be improved by working together university and research council officers in the United Kingdom. The Researchfish, a research collection, and its interaction with the CRIS systems of several universities in the United Kingdom is proposed as a use case.

CRIS and Institutional Repositories

An important issue is the relationship between CRIS and institutional repositories, as complementary systems (Melero, 2020). In fact, this is a basic issue, since CRIS should feed and reflect the content of the repository as far as the university research task is concerned. However, for this to happen, the first premise is that the institutional repository contains all the scientific production of the institution and is the single and updated source of this information (Rodríguez Terán, 2015). This unfortunately is not always the case, resulting in complicating both the implementation and the existence of CRIS.

Although there are examples of all kinds in which either the CRIS is interrelated with the institutional repository and others even in which the CRIS is itself the one that manages the repository itself, as would be desirable, there are still cases such as the one of the university to which it applies, where the absence of an institutional repository, complicates the implementation of the CRIS system. Faced with this problem, which is quite common in the university environment, the CRIS system can centralize the information on the university's scientific research, especially when this is disseminated in different sources. This is not an easy task and requires an extra effort of prior integration of the information in a common repository or database.

This problem will be an additional contribution and benefit of the present system, serving not only as an information visualization system, but also for the integration and management of this information once it is integrated.

Regarding the integration of information from heterogeneous sources, it is going to propose a solution, but it is a latent concern in the design of CRIS systems, as it is also shown in (Azeroual, 2019), where specific ETL (Extract, Transform, Load) techniques are proposed to achieve this integration.

CRIS and its Standard Data Formats

Another transcendental issue in relation to CRIS systems is the format used as a labeling and information exchange system. In this sense, if a common format can be used between institutions, the advantages and benefits would be innumerable and the functionalities of CRIS would be clearly reinforced. It could facilitate the exchange, search and comparison of information and scientific quality between institutions, as well as facilitate the justification of projects and research grants from institutions outside the institution itself.

In this sense, CERIF (euroCRIS, 2021), the common European format for research information promoted by euroCRIS, is probably the most recognized and explicitly created format for CRIS systems at the European level. CERIF is shown as a standard that can help CRIS systems (Sousa Pinto, 2014).

euroCRIS (European Organization for International Research Information) (euroCRIS, 2020) is a European organization created in 2002 that promotes the creation and use of international standard data formats for research information management, such as CERIF. Because of all these characteristics and because it is considered a de facto standard, it is desired to base on this format with some extensions to adapt it to the present needs and make it more complete.

CERIF is not the only standard for CRIS systems. There are other formats such as the model used by DSpace-CRIS (Lyrasis, 2021). This tool uses the institutional repository as its base component, but in this case this institutional repository must first be built by integrating data from different sources. The tool facilitates interoperability with other infrastructures such as OpenAIRE (Open Access Infrastructure for Research in Europe), especially in its latest version where its interoperability is facilitated precisely through CERIF (OpenAIRE, 2020). OpenAIRE (OpenAIRE, 2021) is an Open Access Infrastructure for research to promote open science by aligning European policies and linking researchers, research and data. The CERIF OpenAIRE profile allows to collect and import metadata from CRIS systems, so it will also facilitate the integration with the CRIS (OpenAIRE, 2017).

On the other hand, the proposal is also indirectly aligned with the Hercules project: "Semantics of University Research Data" (Universidad de Murcia, s.f.), promoted by the CRUE in Spain precisely to promote the open university and the use of common semantic standards or formats when publicizing research work in Spanish universities. In this sense, one of the members of the project who has worked

as a reviewer in the alternatives proposed as a standard (Universidad de Murcia, 2021), has contributed to this project looking for a common semantic language for the exchange of information on scientific production between universities. It is also provided to Hercules project, the experience of part of the team in the open publication of information related to the university context (Fermoso García, 2018).

Noting that the proposals assessed in Hercules also took CERIF as a reference and starting point and, as is this case, extended and improved it to complete it. Moreover, the type of information and use cases that gave rise to this extension and improvement of the initial model have coincided with some of those of this proposal. Based on the experience, it has even been made contributions to the alternatives to be evaluated in the Hercules project that have been considered and transferred to the Hercules semantic model that is being developed to become a standard promoted by the CRUE at Spanish national level.

MAIN FOCUS OF THE ARTICLE

Issues and Problems to Solve

Frequently, information about the university scientific production (projects, publications, patents, academic works, ...) and its agents, researchers or TRS staff, organized in research groups, are disseminate in different sources in the university environment. Therefore, it would be necssary to integrate this information before using it, when this information is not collected in an institutional repository or single data system, as is the case and problem of many universities.

On the other hand, it would be very useful not only data can be viewed, such as it is the main goal of a CRIS, but also download and reused them if user desired it, for instant to develop added value services.

Taking into the account the above, it is developed an information system that by integrating data from different sources and applying semantic technologies, allows to publish and share with society the scientific production generated in the university environment.

To solve this problem, the purpose is not only a data visualization and management system, but also a system for integrating data from heterogeneous sources among which research information is disseminated in the university: institutional and library repositories, web pages, intranets and other databases from different departments and services in university.

On the other hand, the system is also implemented as an open access system. Data can not only be viewed, but also downloaded and reused. This is achieved thanks to the use of semantic technologies and furthermore, thanks also to the use of formats based on European standards in the field, such as CERIF, the Common European Research Information Format (euroCRIS, 2021) promoted by euroCRIS (euroCRIS, 2020), which extends the performance even to interoperability with other systems that also share these formats recognized at European and international level.

The system is developed by and for a specific university environment, but with the aim of being reusable in any other university environment, thanks to the use of formats based on international standards in the area.

METHODOLOGY AND STRUCTURE OF THE CRIS PROPOSAL

In the previous point it has been made an overview of the situation and problems of CRIS in the current University and contextualized the reasons for the response given in the present work.

The developed CRIS system is presented as a point of public access to information and dissemination of the production and research facet of the University.

In practice, this implies not only disseminating information related to the University's research activity, which is the main task of a CRIS, but also in this case, being able to collect and integrate it into the same system from different sources and become a single point of access to this activity.

On the other hand, the information already integrated in a single repository or single data system must be visualized and offered in the most efficient way possible. That is, it is also desired that the information can be shared and, where appropriate, even reused in order to be recognized as a true open data system. This added value is achieved by using formats that are recognized as de facto standards, as well as by using technologies such as semantics technologies.

Based on these premises, the system proposal can be considered from the software point of view, organized into three subsystems: information integration and management, visualization, and advanced query and data download. In turn, all these subsystems are supported by a data system that is actually based on two complementary models.

It will now describe the requirements of each of these three subsystems or modules, as well as the characteristics of the dual data model on which they are based.

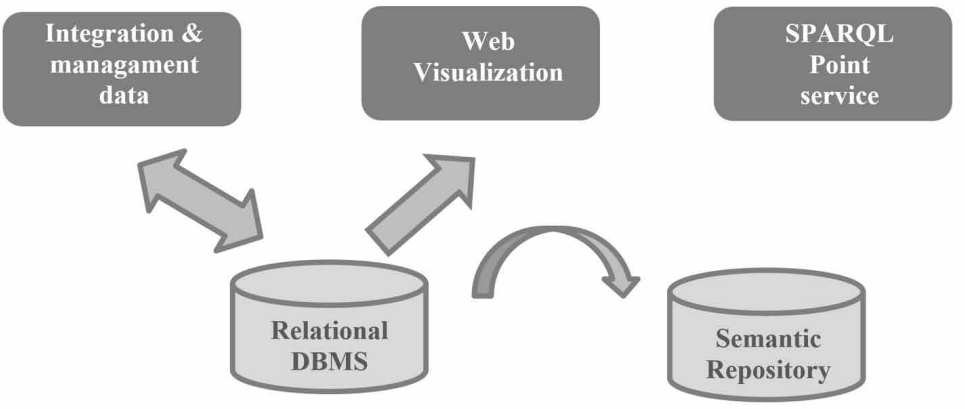
System Software Architecture

The implemented system is organized in three modules. The three modules are interrelated and share a common repository supported by two data models, one relational and the other semantic, which complement each other.

Specifically, the first module is responsible for integrating and maintaining the data on the university's research work in a repository based on a relational database. The second module, the web visualization system, shows the previously integrated information on scientific production: publications, projects, patents, theses and academic works; as well as its agents: TSRs and research groups, organized in a web system, and even including statistics on this production. This module also provides access to the third module, the semantic query service. The third module allows advanced queries to be made in semantic format and to display and/or download the results obtained in different formats. To provide this service, the data from the relational model are previously converted into a semantic format or own ontology. Thanks to this, data are already available to be queried from the SPARQL service.

Figure 1 schematically represents the modular organization or architecture of the system and the repositories that serve as a link between them.

Figure 1. Software architecture of OpenUPSA CRIS



Data Integration and Management Module

As mentioned above, and in the absence of a single data system on the one hand, and on the other, given that the data on research work at the university are scattered across different sources, it is necessary to design a system that allows the data from these sources to be integrated into a single repository. This repository will be supported by a relational database model, which it will be describe later, and which will serve to feed the rest of the system.

The CRIS system will display information on the research staff or TRS, on the research groups in which they are organised, as well as on their scientific output. The research results include publications, projects, patents and academic works such as theses or final degree projects and final master degree projects.

In practice, this information on results is stored in different formats (Leiva-Mederos, 2017). Specifically, the information on the PDI will be collected from the Standardized Curriculum Vitae promoted by the Spanish Ministry of Science and Innovation (FECYT, 2021), known as CVN (Standard Vitae Curricula), of each researcher. The CVNs are stored in PDF files. From the CVN, information is gathered, among others, from projects and publications of the TRS. In the case of publications, the information will also be completed, if necessary, with external sources such as the bibliographic database Scopus (Elsevier, 2021). The information retrieved from the Scopus API (Elsevier developers, 2021) is obtained in JSON format.

System is also prepared for the incorporation of files in CSV format with information on the university's researchers obtained from various sources and provided by the university library.

The information on research groups has to be obtained from the web page where the university publishes information about them. Therefore, in this case information comes from a source in HTML format.

Finally, information on academic work conducted by the teaching and research staff, such as theses, master's or bachelor's degree dissertations, articles in university journals or other publications also from the university and stored in the institutional repository, is obtained from the API of the library's SUMMA repository (Universidad Pontificia de Salamanca, 2021), which allows data on these works to be obtained from the university library repository. This information is obtained in JSON format.

Having analyzed the different sources and formats from which to obtain the information to be integrated and published in the system, and in order to achieve integration, it will be necessary, on the one hand, to propose a common data model associated with the repository that will serve as a single data system. This repository will be supported by a relational model or database. On the other hand, it will be necessary to design and implement tools or analyzers to extract and convert or transform the information from its original sources and formats to the relational model. All these converters to the relational model: PDF parser, CSV parser, Web parser (HTML), JSON parser from the response of different APIs, are one of the fundamental contributions of the project.

The information or fields to be collected from the data source will be mapped and matched with the database fields to which this information must be migrated. The method of accessing each data source and storing its data in the relational database that will act as a repository or single data system, will depend on the type of source and format. In all cases, Java will be the programming language used for the implementation of each parser. In the case of PDF, CSV and HTML sources, the collection of information is done in a more textual way. When data is collected from APIs that return their results in JSON, the structure provided by this format is exploited to facilitate information retrieval.

The logic and method of extracting information from some of these formats can be considered as a model to be reused in other environments, especially the CVN, as this is the standard curriculum format for the TRS in all Spanish universities. The CVN analyzer is therefore a fully exportable tool.

The data collected depends on each source and then it will be visualized and also published in open formats that facilitate its download and reuse.

For example, as part of the initial upload to the system, the CVN collects the personal data of the teaching and research staff, as well as their publications and projects. Afterwards, the updates will be carried out automatically and/or programmed from the CRIS intranet by the CRIS administrators.

Using the Scopus API and the researcher's Scopus ID, information on their publications are enriched. This information can be used to extend the information obtained for the same publication from the CVN of the TRS, such as the abstract.

Finally, information on the academic work published and directed by researchers within the university is extracted mainly through the API provided by the SUMMA repository of the university library. The SUMMA query service implemented through its API, returns data on theses, students final works, journal articles and publications of the university authors, returning the results in JSON format.

Up to this point, it has been described how data integration is carried out and the tools or parsers from different formats developed for this purpose. However, integration is not the only functionality of this module. Once the data has been integrated and stored in the relational database that acts as a common repository, it can be also updated, managed, maintained or deleted if desired, by CRIS system in order to always maintain data updated.

It is also important to note that this data integration and management module described above is not public, but only for internal use and maintenance by authorized library staff and/or other actors.

In the results section, the interface that allows carrying out some of the tasks described in this module will be shown later in this work, as well as the structure of the database that supports the repository or single data system of the CRIS is also detailed.

Web Display Module

The data integration and management module just described, is for internal use only and has been developed to make up for the current lack of a single data system at the university for the research production of the institution. However, the fundamental aim of a CRIS is to make this information known externally. To disseminate the scientific production and research work carried out by the University's teaching and research staff, usually organized in turn into research groups.

To meet this objective, a web information system has been developed to show, from the information previously stored in the new repository and by means of SQL queries to it, the data on research groups, with their description, lines of work and researchers that make them up. For each researcher or TRS in particular is displayed his/her data and also link to their publications, projects or academic work, which on the other hand can also be viewed independently. About publications, if applicable, it can be even displayed their abstracts if they are published in Scopus, thus also providing links to external sources.

In addition, it also offers the possibility of visualizing different statistics on this scientific production by dates or categories or impact for example, which can even contribute to the university's decision making process, as well as in any case, provide external indications on the quality of the institution's scientific production.

The web system to provide all these functionalities is organized in different sections, and its interface is shown in the next results section. Finally, this system also provides access to the last of the modules, the advanced query and download module.

Advanced Query Module

The SPARQL Point module or service, together with the integration module, is one of the most innovative and differentiating contributions compared to a traditional CRIS system, which allows us to consider the present CRIS system as a high-performance advanced CRIS.

Thanks to this module and the technologies and formats that support it, it can be considered that the information is really published openly with the highest performance. The previous module, the Web system, allows the information to be made public or visualized by displaying it on the Web, but it does not allow this information to be reused or exported, for example to other contexts, and for

this reason it cannot be considered as an open access system, which is what it should be aspired to in order to provide a better service to the user and data transfer.

In turn, when it is published in open allowing not only to view, but also to download data, this is made at different levels if it is considered the scale established by Tim Berners Lee (Berners Lee, 2014), the creator of the Web. This scale has 5 levels so that at the highest level, open publishing allows data to be more reusable and even linkable, by sharing it openly. This higher level of open data quality should be aspired to, and this is what it is aimed for in the present system, as a service in addition to the basic CRIS system.

In this sense, this third module can also be defined as what is commonly called a SPARQL Point service. From it, the user can make any particular queries to the system, in semantic format or SPARQL language (Semantic Query Language) (w3c, 2013) . It also allows not only to visualize the results obtained when making the query, but also to download them in different formats such as JSON or RDF, according to the user's preference, for later reuse and creation of value-added services from them if necessary, or for example also to link or exchange them with other CRIS systems, especially if they use the same standards.

To carry out all this task, a semantic data model or ontology has been created, which is the basis of the data repository that is maintained in parallel and fed by the central repository created by the first subsystem and based on a relational model. From the relational model the data is converted to the semantic model. From there, the data can be queried by means of semantic queries in SPARQL language. Thanks to the SPARQL Point service designed, the results obtained from the query can also be downloaded. All this is possible thanks to the use of semantic technologies and the use of ontologies. Specifically, an own ontology has been designed, but at the same time based on the European CERIF standard, as will be detailed later on. Based on this semantic model, an import or data conversion service from the relational model to the ontological model has been designed, on the one hand, and a semantic query service in the SPARQL language, on the other. Finally, processes have also been implemented to allow the download of data obtained in different formats (JSON and RDF). The results section will show the interface that allows all this to be achieved, and in the following section on the data model, the structure of the semantic or ontological model will be detailed.

Data Models

As mentioned indicated in the system architecture, the CRIS is based on two data models that complement each other.

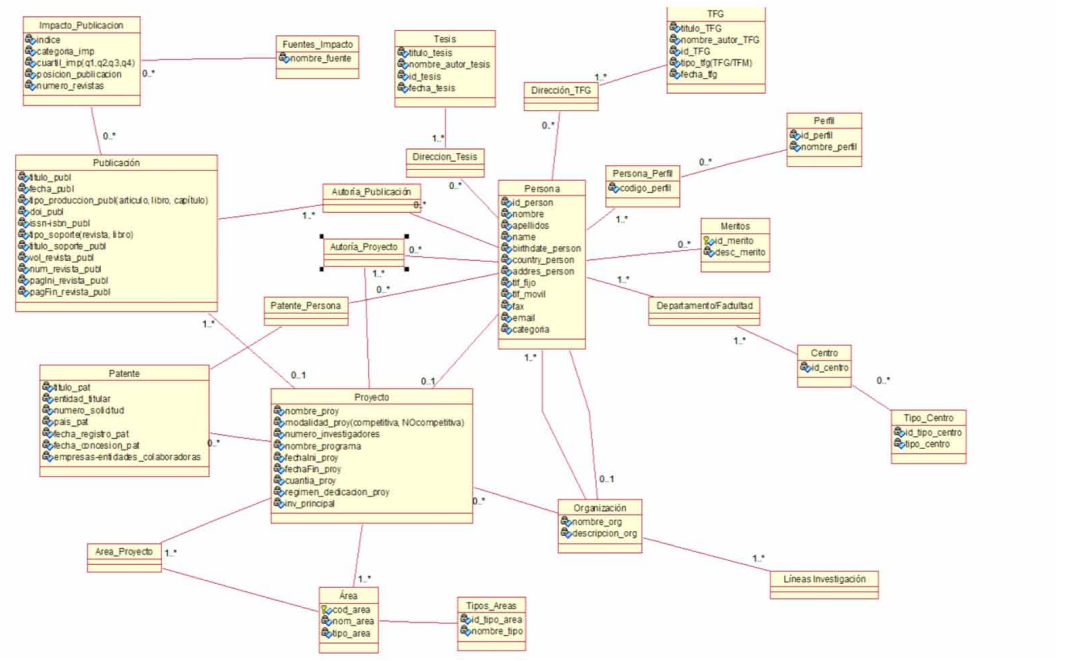
On the one hand, it has been designed a single data repository based on a relational model. On the other hand, it has also been designed a semantic model that feeds on the first one and allows us to turn the proposal into a true open data system that is interoperable and revalued by being based on a European standard such as CERIF. The structure of the entities of these models, their attributes and interrelationships in both models, derive from the specification of requirements or use cases or services to be offered by modules of the system previously described.

Relational Data Model

The database is based on 8 main entities and other secondary entities derived from the interrelations and standardization of the previous ones: PDI, Organization, Publication, Project, Patent, TFG, TFM and Thesis.

The structure of the relational database with its fields and relations is shown in *Figure 2*. This model is implemented on a MySQL database management system.

Figure 2. OpenUPSA Relational database model



Ontology

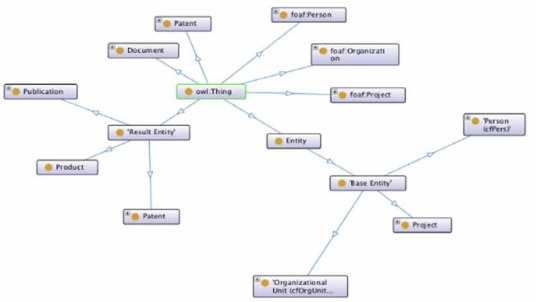
The semantic model, whose data comes from the relational repository, is based on an ontology called OpenUPSA.

This ontology is based on another existing ontology such as CERIF, which is recognized as a standard format for European CRIS systems promoted by euroCRIS. It is precisely for this reason that it has been taken as a reference, although modifying and extending it to adapt it to the special needs.

The advantage of adopting this model is that system is compatible with other CRIS at international level, for example when exchanging or comparing data. Nevertheless, the CERIF model has certain limitations for some of the use cases that the system wishes to provide. Therefore, it has been designed a semantic model based on CERIF, but updated in order to adapt to the specific needs.

The customized CERIF-based ontology model, which it has been named OpenUPSA, is shown in *Figure 3*.

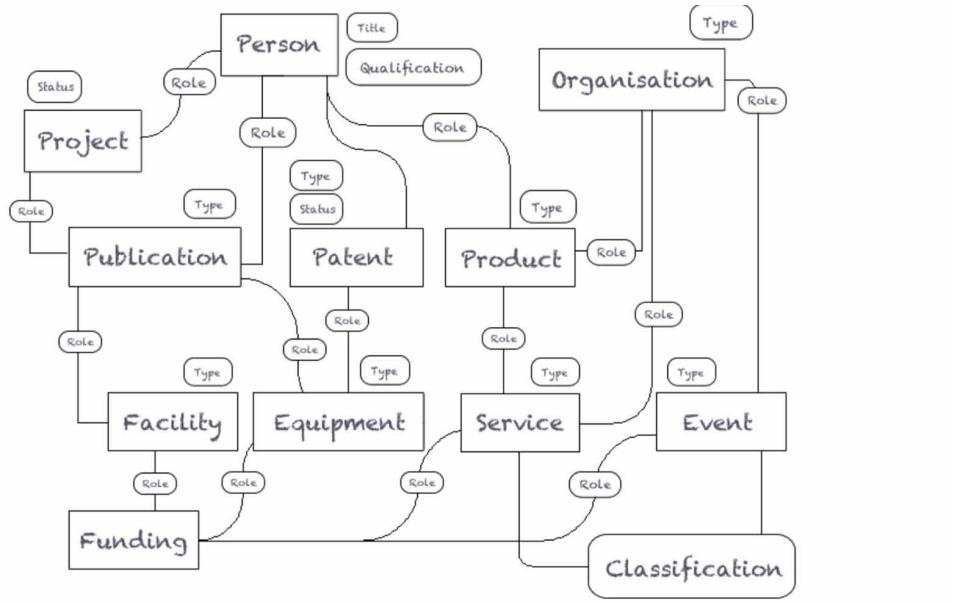
Figure 3. OpenUPSA ontology model



This model follows the CERIF model without adding new classes, but adding in some of the classes new properties necessary for the CRIS model.

The three main entities are Person, Organization and Project, classified as base entities. Another important entity to highlight is Publication. All four classes are CERIF-specific, as shown in *Figure 4*.

Figure 4. Visualisation of selected CERIF 2000 data model entities, roles, statuses and types Source: (https://www.eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html)



For each of the classes, all the necessary characteristics provided by CERIF have been used, which will be adequately indicated, and characteristics (Properties) have been added by OpenUPSA to adequately describe the individuals. Annotation Properties have been used for their properties. For the relationships between individuals, Object Properties have been used.

Specifically, some properties have been added to the CERIF classes that are necessary to represent the information to be shown in the system, such as the following:

. Person: ORCID as identifier for the researcher.
. Project: budget or amount, main researcher or number of participants.
. OrganizatioN: description of the organization or research group.
. Publication: DOI identifier, ISSN-IBN, start and end pages of the publication, title and type of publication and data source of publication, ...

Figure 5 shows more detail about the ontology and its structure of classes and relationships.

Figure 5. OpenUPSA ontology classes structure

RESULTS

Having described the organization of the system made up of three subsystems, their interfaces are going to be shown in this section.

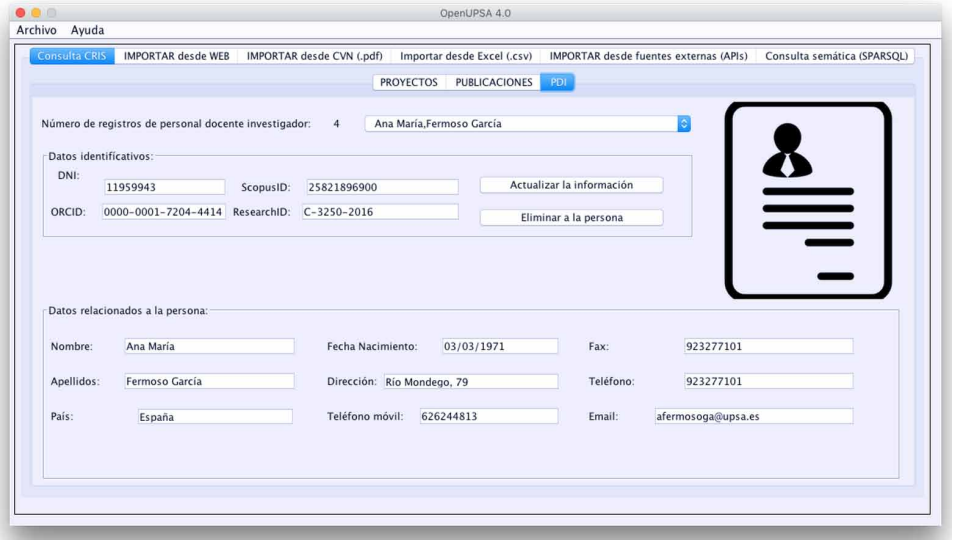
Data Integration and Management Subsystem

The purpose of this module is to integrate in a single repository based on a relational database, data on scientific production and its agents in the university.

This subsystem is only accessible internally from the university and by authorised users, in principle university library staff, as this department is mainly responsible for the custody and dissemination of the institution's scientific production.

Once inside, the system provides the appropriate interface for selecting and integrating each selected source into the repository according to its format. In addition to displaying the data on the repository's main entities, this option also allows you to update their information if necessary. It is therefore the section corresponding to the data management service in addition to the integration service. *Figure 6* shows, for example, the management and visualisation of data from a TRS.

Figure 6. Teaching Research Staff (TRS) view and management detail



On the other hand, there is an available option to import for each different type of source. Below it is shown, as an example, the appearance of the data import process for some of them, such as the import of the data of the TRS from their CVN, *Figure 7*, or from the Web for the research groups *Figure 8*.

Figure 7. Import CVN view

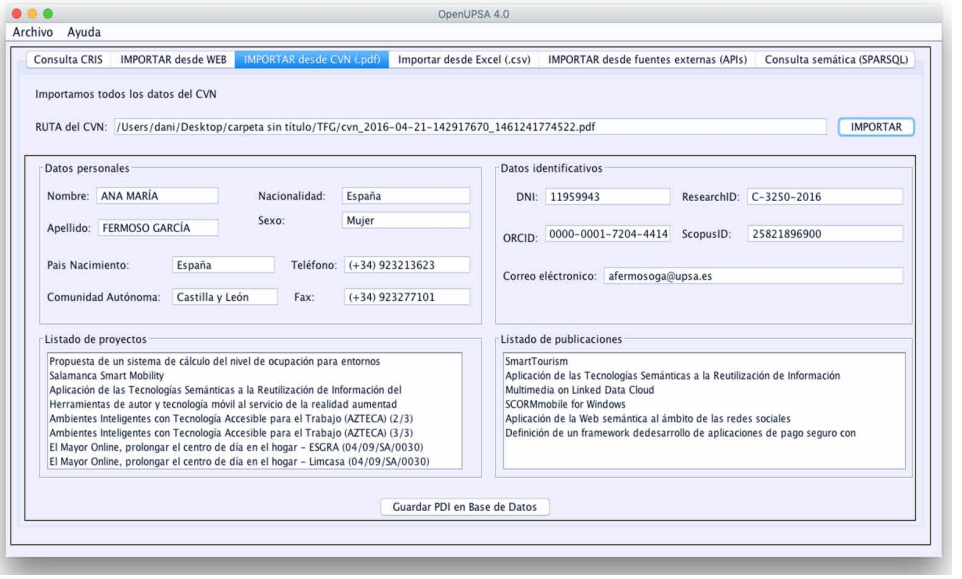
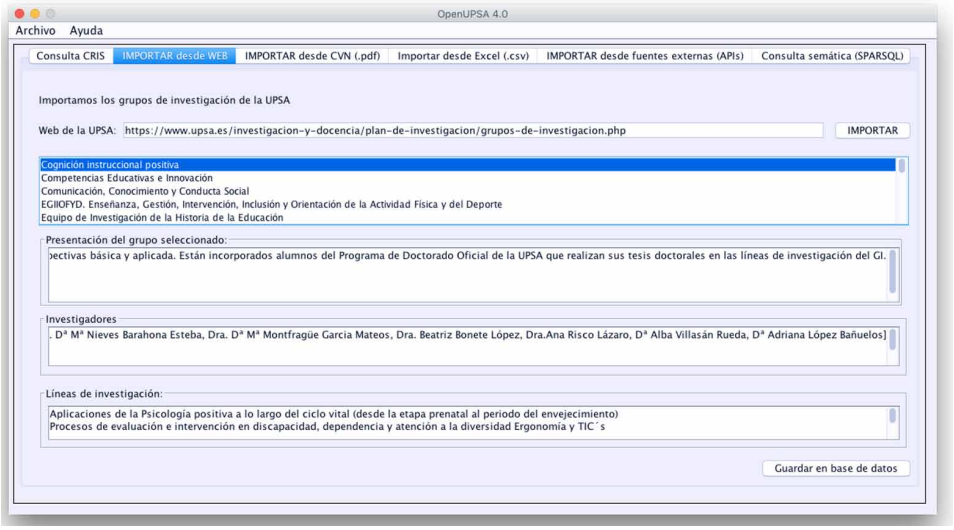


Figure 8. Import research group information view



Web Subsystem for Public Viewing of the University's Research Facet

This module is the web system organized by sections, accessible by any user inside or outside the university and from which the main entities related to research can be viewed. The information displayed will depend on the type of entity selected and between them can be navigated in a linked way, so it is possible to reach the same information from different points. Thus, from a group, for example, it can be accessed the TRSs that make it up. Likewise, publications or projects can be accessed in particular, or through the profile of the TRS associated as author of that publication or project.

Figure 9 shows the appearance of the main website. From this subsystem there is also access to the advanced queries service or SPARQL Point.

Figure 9. Home page of the OpenUPSA CRIS system

The main page is the gateway to the CRIS model and is presented in a simple way, categorizing the information available in different sections.

The TRS (PDI) option shows the list of researchers (*Figure 10*) and by clicking on a particular researcher, HIS/HER details are displayed (*Figure 11*) (name of the researcher, e-mail address, ORCID and the research group to which he/she belongs, which can also be accessed from the link to their group). In turn, each researcher can access their scientific production: publications, projects or academic work conducted. All these research results can also be accessed from the corresponding specific option.

Figure 10. List of researchers view

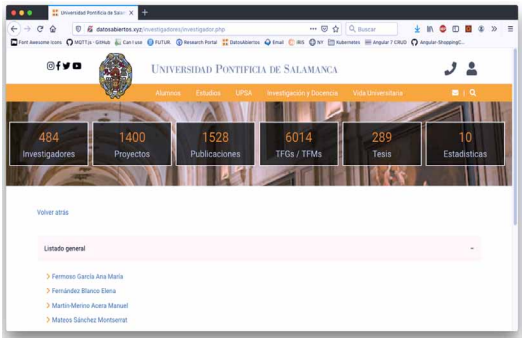
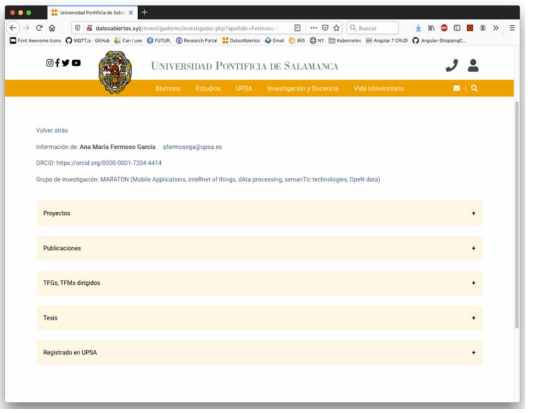


Figure 11. Researcher information view

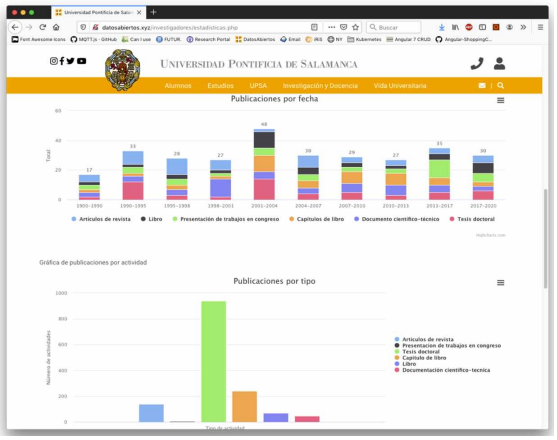


Finally, and particularly interesting, is the Statistics option. As its name suggests, it shows, in graphic format, various statistics that can be used as a showcase when it comes to publicising the quality of scientific production, as well as for internal decision-making with a view to continuing to promote its improvement.

Multiple comparative statistics can be displayed based on different criteria and all of them can be exported to formats such as pdf, svg, png, etc.

As an example, *Figure 12* shows a comparison of the number of publications by type and date.

Figure 12. Publications by type

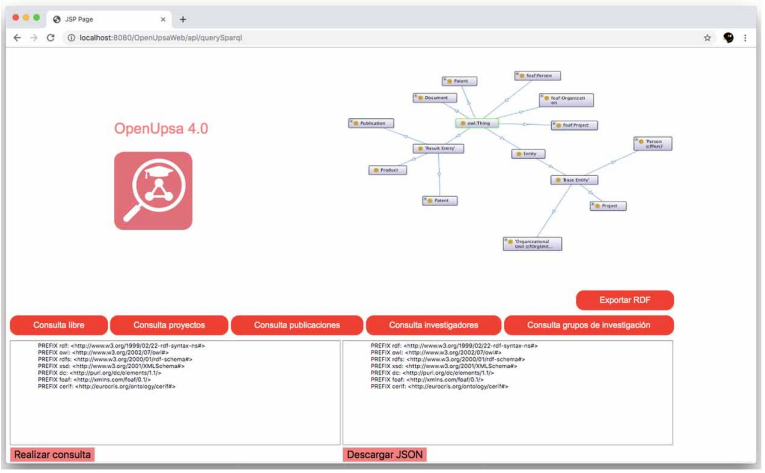


Advanced Semantic Query Subsystem

Finally, it is shown the advanced query service or SPARQL Point (*Figure 13*). Through it, they are performed specific queries to the repository and according to the model shown with the structure of the ontology at the top of the screen, to facilitate the user's task. Likewise, with the same objective, the user is given the option of executing the most usual predefined queries, or writing by his own, in whose case the. "PREFIX" structure to be included in any query to the system is already predefined.

The query results are displayed on screen, but can also be downloaded in JSON or RDF format for reuse by the user. This represents one of the most advanced features as a CRIS system, as well as facilitating the reuse, sharing and exchange of information with other CRIS and organizations, if desired. The queries use the SPARQL language, so it will be the more advanced users with a good command of this language who will be able to make the most of this advanced service of the system.

Figure 13. SPARQL Point service view



CONCLUSION

Having analyzed the system, its features and appearance, it is concluded that its characteristics go beyond what a CRIS system is by incorporating additional advanced features.

On the one hand, the CRIS offers an alternative and thus solves the problem of the absence of a single repository system for research, an important and fairly common problem in the university environment, which it is still working to solve. The integration system is valuable in itself, but also for the format conversion tools that it provides in the form of parsers. Particularly interesting is the parser that makes it possible to extract information about the researcher and his or her production from his or her CVN. This CVN or standardized curriculum vitae is the one recognized by the state university system, and therefore reusable as a tool in many other possible contexts and/or to create value-added services based on the information extracted from any researcher at any university from their CVN. Equally interesting is the fact that the system has also implemented a service that allows the data of a publication to be enriched from external sources, specifically with the information that may appear in such renowned bibliographic databases as Scopus.

The second interesting contribution is the possibility of advanced querying and downloading of data based both on the use of semantic technologies and on the semantic format itself, based on a standard recognized at European level. This makes it possible to internationalize the proposal and also to extend its model to other CRIS systems, as well as to be recognized as a true open access system, by allowing the information to be downloaded and reused by users outside the institution.

Another important contribution is that based on their experience, the group of the project has collaborated in the construction of a national standard for CRIS systems in Spanish universities, promoted by the CRUE through the Hercules project for an open university, currently still under development, but already quite advanced, and in which the team continues to participate and contribute.

However, the system implemented is a pilot that can be further improved. For example, expanding the number and type of statistics with new use cases, generating new data representation graphs. Likewise, and in order to continue enriching the data with external sources, not only linking with Scopus, but also with other prestigious external sources such as WOS, ORCID, Scholar or others, provided that these sources, as is the case with Scopus, make an API available to us for this purpose.

On the other hand, and taking into account that the university has the dual responsibility of teaching and research, as already mentioned, another line of research and improvement would be that the system designed could also contribute to the more academic facet. An analysis of the contribution of the semantic web to educational systems is shown in (Hu, 2022). The paper shows different examples of how the semantic web helps educational systems to respond to the needs of their users. Students must acquire knowledge from different sources. In this sense, ontological resources to help students should be linked in the first instance with the scientific production of their university itself, linking the ontological system of the most academic educational system with the scientific production or CRIS system, thus recommending to the user, if necessary and in the first instance, research resources provided by the university itself and its authorities, for the improvement and enrichment of their learning.

Along the same lines, a system for recommending courses to students on e-learning platforms is proposed in (George, 2021). The same system could also be used to recommend library resources published in the university CRIS to students, after defining their needs and profile.]

Another different extension of the project could be not only providing access to the resources of the institutional repository of the university library, but also on step more by providing access to the own content, that is, to the text of the resource. Continuing with the use of semantic technologies and ontologies, in (Stylianou, 2022) is proposed a framework called Doc2KG that allows to transform textual documents in knowledge graphs to classify an navigate through the content of the text document. Despite of the fact that this framework is designed to transform into knowledge graphs mainly related

to government documents, nevertheless if could also be adapted to a new ontology for classification by content of the document resources presents in the library repository of University.

A second possibility to classify by their own content the documents published by the CRIS system is in (Sharaff, 2021) where it is proposed an efficient framework for the generation of better topic coherence, where term frequency-inverse document frequency (TF-IDF) and parsimonious language model (PLM) are used for the information retrieval task and based on it, the documents cab be olso be classified by its main topic, like in the previous mentioned work.

REFERENCES

Abadal, E. (2019). *Los sistemas de gestión de la investigación (CRIS): ¿cómo se utilizan?* http: //www.ub.edu/blokdebid/es/content/los-sistemas-de -gestion-de-la-investigacion-cris-como- se-utilizan

Anna Clements, G. R. (2017). Let's Talk – Interoperability between University CRIS/IR and Researchfish: A Case Study from the UK. *Procedia Computer Science*, *106*, 220–231. doi:10.1016/j.procs.2017.03.019

Azeroual, O., Saake, G., & Abuosba, M. (2019). ETL Best Practices for Data Quality Checks in RIS Databases. *Informatics (MDPI)*, *6*(1), 10. doi:10.3390/informatics6010010

Berners Lee, T. (2014). *5 estrellas datos abiertos*. Obtenido. https://5stardata.info/es/

Bryant, R., Clements, A., Castro, P. d., Cantrell, J., Dortmund, A., Fransen, J., & Mennielli, M. (2018). Practices and patterns in Research Information Management: findings from a global survey. *OCLC Research*, *18*. doi:10.25333/BGFG-D241

Ebert, B. (2014). Using a CRIS to reduce workload and increase quality for research reporting and university marketing. *Research Information Systems: Integration for Open Access to Scientific Outputs*, 31-48. http://hdl.handle.net/11366/237

Elsevier. (2021). *Scopus. Expertly curated abstract & citation database*. https://www.elsevier.com/solutions/scopus

Elsevier developers. (2021). *Elsevier Scopus APIs*. https://dev.elsevier.com/sc_apis.html

euroCRIS. (2020). *euroCRIS. The international organizarion for research information*. https://eurocris.org/

euroCRIS. (2021). *CERIF in brief*. https://www.eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html

FECYT. (2021). *Editor CVN*. https://cvn.fecyt.es/editor/#HOME

Fermoso García, A. M. (2018). Sistema de Modelado Semántico para catalogación, clasificación, consulta y publicación en abierto de información bibliográfica. *El Profesional de la Información*, *27*(2), 410–418. doi:10.3145/epi.2018.mar.20

George, G. L. (2021). A Personalized Approach to Course Recommendation in Higher Education. *International Journal on Semantic Web and Information Systems*, *17*(2), 100–114. doi: doi:10.4018/IJSWIS.2021040106

Harris, S. & Seaborne, A. (2013). *SPARQL 1.1 Query Language*. https://www.w3.org/TR/sparql11-query/

Hu, B. G. (2022). Evaluation and Comparative Analysis of Semantic Web-Based Strategies for Enhancing Educational System Development. *International Journal on Semantic Web and Information Systems*, *18*(1), 14. doi: doi:10.4018/IJSWIS.302895

Jeffery, K. G. (2006). Supporting the Research Process with a CRIS. *Enabling Interaction and Quality: Beyond the Hanseatic League*, 121-130.

Leiva-Mederos, A. S.-D. (2017). Working framework of semantic interoperability for CRIS with heterogeneous data sources. *The Journal of Documentation*, *73*(3), 481–499.

Lyrasis. (2021). *DSpace CRIS*. https://wiki.lyrasis.org/display/DSPACECRIS/

Malo de Molina, T. B. (2018). *Estado de la cuestión de los CRIS en las universidades españolas.* https://e-archivo.uc3m.es/handle/10016/27681

Melero, R. (2020). *¿CRIS versus IR?* https://www.ub.edu/blokdebid/es/content/cris-versus-ir

OpenAIRE. (2017). *OpenAIRE Guidelines for CRIS Managers*. https://openaire-guidelines-for-cris-managers.readthedocs.io/en/v1.1.1/

OpenAIRE. (2020). *Implementation of the OpenAIRE-CRIS-CERIF Guidelines in DSpace-CRIS*. https://www.openaire.eu/blogs/implementation-of-the-openaire-cris-cerif-guidelines-in-dspace-cris

OpenAIRE. (2021). *OpenAIRE*. https://www.openaire.eu/mission-and-vision

Programmable Web. (2021). *Scopus REST API*. https://www.programmableweb.com/api/scopus-rest-api

Rodríguez Terán, A. (. (2015). *Sistemas de Gestión de la Investigación: aproximación a los CRIS Institucionales*. https://gredos.usal.es/handle/10366/129659

Sharaff, A. D. (2021). Prospecting the Effect of Topic Modeling in Information Retrieval. *International Journal on Semantic Web and Information Systems*, *17*(3), 17. doi: doi:10.4018/IJSWIS.2021070102

Sousa Pinto, C. S. (2014). CERIF – Is the Standard Helping to Improve CRIS? *Procedia Computer Science*, *22*, 80–85. doi:10.1016/j.procs.2014.03.013.

Stylianou, N. V. (2022). Doc2KG: Transforming Document Repositories to Knowledge Graphs. *International Journal on Semantic Web and Information Systems*, *18*(1), 20. doi: doi:10.4018/IJSWIS.295552

Universidad de Murcia. (2021). *Hércules. Proyecto ASIO*. https://www.um.es/web/hercules/proyectos/asio

Universidad de Murcia. (2022). *Hércules*. https://www.um.es/web/hercules/inicio

Universidad Pontificia de Salamanca. (2021). *SUMMA Repositorio Institucional*.

Ana María Fermoso García has a PhD in Computer Science and Computer Engineering from the University of Deusto. She is currently Professor of Software Engineering at the Faculty of Computer Science of the Universidad Pontificia de Salamanca. As a researcher she has institutional recognition, projects and publications of impact. Her lines of work have focused on semantic technologies and the semantic web, for which she has impact publications, projects and participations in conferences related to this area. This topic is also one of the main lines of work and research of the MARATON research group of the UPSA, to which she belongs. More specifically, she is particularly interested in the publication and reuse of open data through semantic formats, such as ontologies, including their design, implementation and exploitation in different fields. In recent work, all of this has been applied, among others, to the university and library environment.

Maria Isabel Manzano García is a Graduate in Documentation from the University of Salamanca. She has a Master's Degree in Digital Documentation and a Master's Degree in Search Engines from the Universitat Pompeu Fabra. In a continuous process of learning and adaptation, she has been a librarian for more than 25 years at the Library of the Pontifical University of Salamanca. After a long period as Deputy Director, she currently holds the position of Director of the Library and Archive Service of the Pontifical University of Salamanca.

Roberto Berjón Gallinas has a PhD. in Computer Science and Computer Engineering from the University of Deusto. He is currently a professor at the Faculty of Computer Science at the Universidad Pontificia de Salamanca. He is a member of the MARATON research group, where he works on research lines related to IoT, data processing and mobile technologies. This research experience is accredited through institutonal recognition, numerous publications indexed in relevant positions in the main international rankings (JCR, SJR), as well as intellectual property registrations derived from the software products developed in different research projects in which he has participated as principal investigator.

Montserrat Mateos Sánchez has a PhD in Computer Science in the area of Languages and Systems from the University of Salamanca. She is currently Professor in Charge of Chair in the Faculty of Computer Science at the Universidad Pontificia de Salamanca. Regarding her research career, she has a six-year research period; on the other hand, she is responsible for the research group MARATON, a group whose main research lines are in the field of information systems, more specifically, in information retrieval, semantic technologies and open data, focusing on the context of mobile technologies and IoT. She has participated as a collaborating researcher and principal investigator in competitive projects related to her lines of research. She is author and co-author of numerous prestigious scientific publications, and has participated as a speaker in several national and international conferences.

María Encarnación Beato Gutiérrez has a PhD in Computer Systems and Computer Engineering from the University of Valladolid. She is currently Professor of Programming Languages in the Faculty of Computer Science at the Pontifical University of Salamanca. She is a member of the MARATON research group, where she works on research lines related to IoT, data processing and mobile technologies. This research experience is accredited through two six-year research periods, numerous publications indexed in relevant positions in the main international rankings (JCR, SJR), as well as intellectual property registrations derived from the software products developed in different research projects in which he has participated as principal investigator.