Building Interoperability in Archival Management Systems, Between Data and Records Long-term Digital Preservation for Joint Research. Challenges in a Fast Changing World[1]

Josep Ballber, Industrial PhD Candidate, Universitat de Lleida
josep.ballber@udl.cat

Pepita Raventós, University Archive Manager (Arxiu i Gestió de Documents), Universitat de Lleida pepita.raventos@udl.cat

Roberto García, Associate Professor, Computer Science, Universitat de Lleida
roberto.garcia@udl.cat

1. INTRODUCTION

This poster aims to promote the research we are doing at the University of Lleida (UdL) in long-term digital preservation. The objective is to find the best way to digitally preserve documents and data within a university context based on real case studies. We have found that to make a good long-term digital preservation so documentary evidence is deposited in digital archiving applications, we must first do good records-keeping management. Our approach is based on real case studies of the university archives and within the framework of the management system of records (MMS) based on UNE ISO 30301:2019. This management system at the UdL has been audited and certified since 2015.

1. RESEARCH DEVELOPMENT

This entails making a description of the object that we want to preserve from its metadata. but current standards such as ISAD(G) have an overly hierarchical and rigid structure and it is what has made research opt to model from the conceptual model of Records in Context with a relational and graph structure, with a more semantic vision based on ontologies. It has also been found that in this process of long-term digital preservation various systems or devices (agents) intervene, and through metadata management interoperability between the different systems and devices (agents) must be facilitated, taking into account the rules that affect the process.[2]

The methodology that has been used is first to use the ability to observe instead of going from technology directly to a solution. The objective is to have the solution well defined beforehand, so the problem or need has been well observed to find the technology that best helps to create it.

Semantic technology has been used while exploring it to analyse how models fit together with Records in Context (RiC) and eArchiving specifications. Also using the microservices architecture to process information (collecting documents and (meta)data...) and creating a common metadata management model that should help use preloaded patterns of metadata schemas. Solutions have also been explored by applying artificial intelligence and *machine learning* to record identification. The work is being developed in three areas of case studies: academic records (Administrative/academic document UdL); teaching guides (digital assets in the field of teaching of the Virtual Campus of the UdL) and didactic notebooks (digitized records of the old Normal School of Lleida, for the field of research for the study of records and data).

This poster focuses on the academic events of the UdL as a case study. It is based on a digital ecosystem of origin with different information systems:

---

[1] Pòster 46 presentat en el marc de la Conferència Anual del Consell Internacional d'Arxius (ICA). Roma, 21 – 23 setembre 2022. https://www.ica.org/es/ica-roma-2022-conferencia

[2] INTERNATIONAL COUNCIL ON ARCHIVES. EXPERT GROUP ON ARCHIVAL DESCRIPTION. "Records in Context. Conceptual Model". Consultation draft v.02. July 2021. https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf [Consulta 18/09/2022]

a) Business management :

1.	Academic Manager (UXXI OCU) > Academic Events Module
2.	Sign Application (@Firma)

b) Archive management :

1.	Documents and Records Management System (GDC from CSUC)
2.	Digital archive. (eArxiu)

The metadata of business applications has been collected in a file (metadades_origen.xlsx) and has been obtained from the PDF of the academic record and the information/metadata contained therein that comes from the two applications mentioned above.

The metadata of document management and digital archive are specified:

- Documents and Records Management System (GdC CSUC) vocabulari de metadades de gestió documental (fons, sèrie, uds) que compleix amb l'esquema e-EMGDE (Metadatos para la Gestión del Documento Electrónico de ENI). metadades_desti_GdC.xlsx
- Digital Archive (eArxiu CSUC) metadades de preservació digital que segueix l'esquema de metadades PREMIS (*Preservation Metadata Maintenance Activity*) metadades_desti_eArxiu.xlsx

The example is an academic record of the degree in social work: "GRAU EN TREBALL SOCIAL: 2020-21(UNI) Assign. 100948 - PSICOLOGIA SOCIAL Grup 1 Ordre 0 .pdf_firmado-1.pdf".

The example of an academic record follows the documentary model fond, series and record) and uses a hierarchical classification plan QdC (Quadre de Classificació), the table of common classification of Catalan public universities, and therefore there are inheritance relationships:

- F(fond) -> Fons de la Universitat de Lleida
- CL(class) -> M1000 Ordenació de la Docència (teaching sorting)
- SCL (Subclass) -> M1024 Avaluació acadèmica i docent (Academic and teaching evaluation)
  - D4 (level 4 of the hierarchical leaderboard) -> M1025 Avaluació dels estudants (Student assessment)
  - D5 (level 5 hierarchical leaderboard) -> M1030 Actes de qualificacions (Grading minutes)
- S (serie) -> M1030N4195 GRAU EN TREBALL SOCIAL (Degree in Social Work)
- UDS (Simple documentary unit) -> 2020-21(UNI) Assign. 100948 - PSICOLOGIA SOCIAL Grup 1 (Social Psychology)

The documentary series is composed of the classification code (M1030 Grading Acts) and the nominative subdivision (N4195 GRAU EN TREBALL SOCIAL). The latter corresponds to the name of the teaching plan (study where this plan is associated).

The identified metadata that use UXXI i @firma should be moved to GDC and eArxiu:

- The archivist manually through the graphical interface of the application.
- With the automation of the process through the API.
- Automatically generated by the GdC when making the capture or deposit (such as timestamps, audit, access, disposition, etc.)

This record-keeping model is hierarchical and this is the structure that reproduces the application of EDRMS system and eArxiu.

The minutes are found by the EDRMS (GdC), the study plans to make the load automatically once signed by the teacher. The example of a record has not yet been transferred to the digital archive (eArxiu) but it is planned to be done with the metadata reported to create the archive information package (PIA).
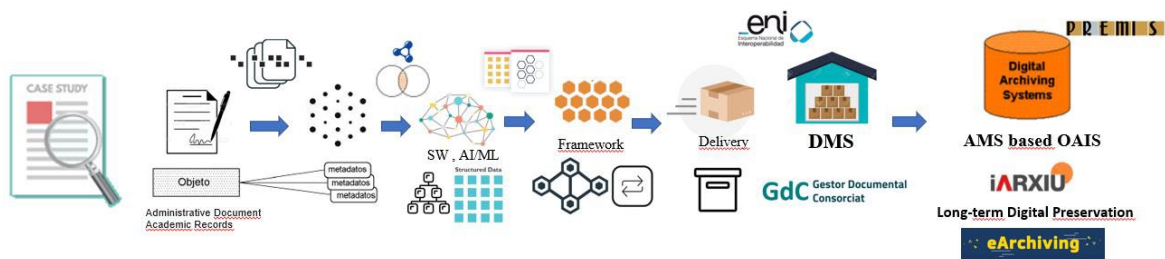

Figura 1. Procés definit per la preservació digital de les actes acadèmiques

STUDY OF THE RECORDKEEPING MODEL

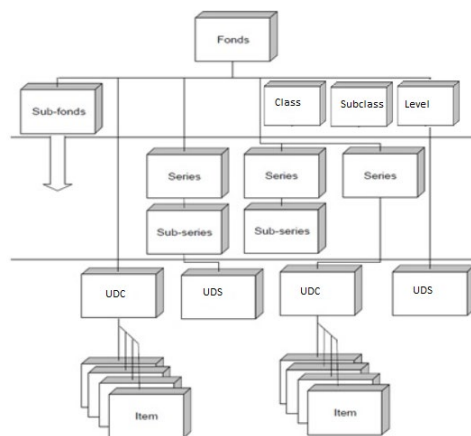The multilevel model according to ISAD (G) is shown in Figure 1.

Figure 1. Multilevel rendering

We start from a **Holding/Fund,** the set of records generated by an organism from the exercise of its activities. This is divided into different classes until reaching the **series,** the grouping or set of records that respond to the same function or administrative activity. And finally, the **Documentary Unit,** the document/record (ric:Record) or set of documents/recirds (ric:RecordSet) that is considered an independent unit in any process or phase of records keeping process. The documentary unit can be simple (documentary piece) or composed (regulated or unregulated file, set of documents related by oneself a subject or business) and that represents the smallest archival unit. They should always be part of a series.

This multilevel model is proving that it is running out because it does not allow to take into account the different contexts through which the document passes and therefore remains undocumented.

At this point, the Multientity Documentary Model can help, according to UNE ISO 23081-2:2021, shown in Figure 2.
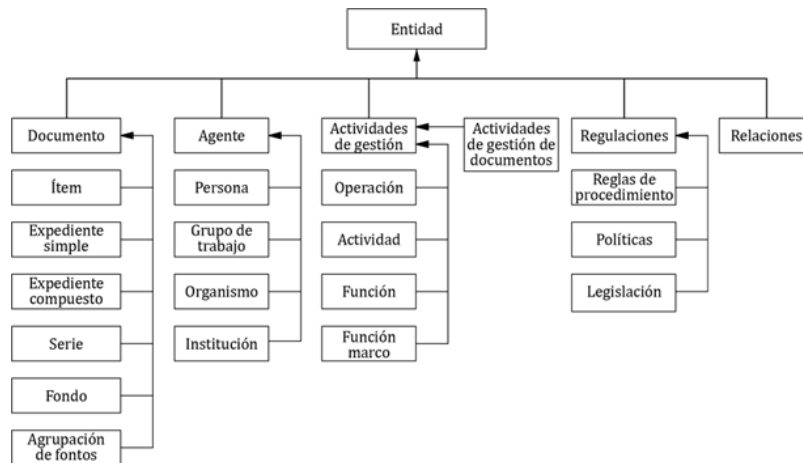
Figure 2. Multientity representation

This model has served the ICA's Archival Description Working Group to write Records in Context and give arguments as to why this evolution has evolved. The result is shown in Figure 3. The hierarchy of entities according to R&C is:
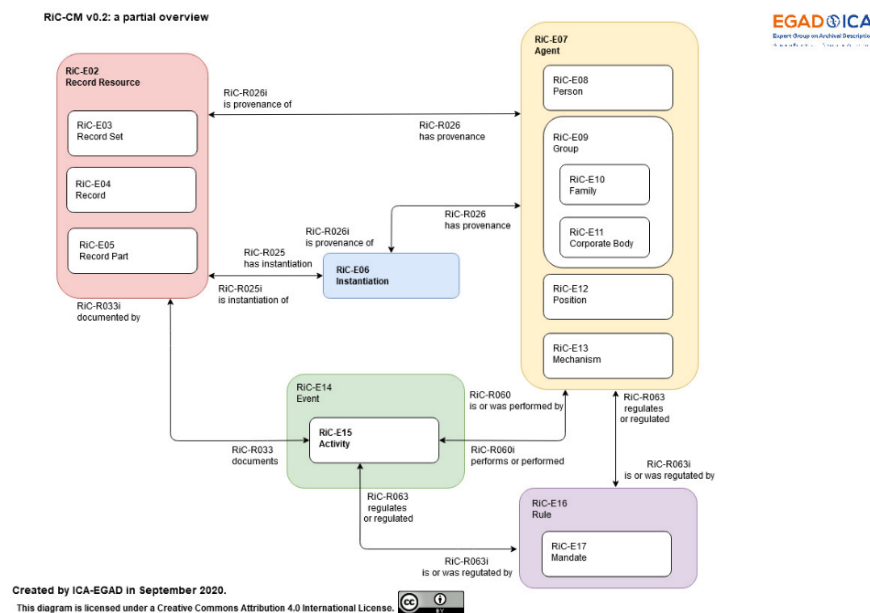

Figure 3: Representation of the conceptual model of R&C (Records in Context)

It is studied whether the model becomes multidimensional and allows compliance with the multientity model defined by ISO 23081.
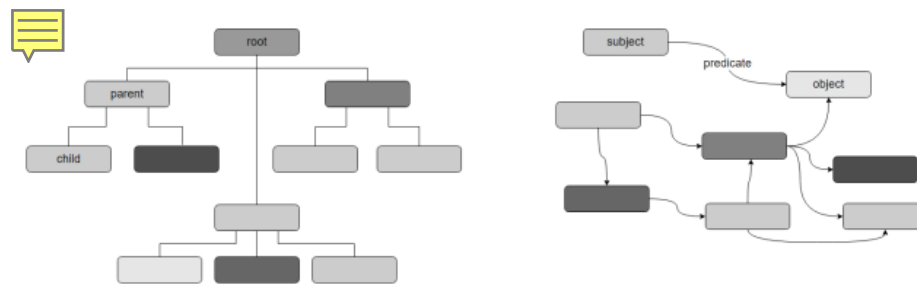
Figure 2. Left: Representation of data in a hierarchical structure like XML or other markup-language. Right: Representation of data in triples that results in a graph-structure.

To carry out good document management we must make use of archival classification, an operation that consists of establishing the categories and groups that reflect the hierarchical structure of the records keeping model, the set of documents generated by an organism from the exercise of its activities.

To classify documents/records, a tool or instrument called QdC (Classification Chart) is used, as shown in Figure 5. It is a document that gives you the indications/guidelines on how to classify records/documents. The Classification Plan of (QdC-UdL) University of Lleida since 2012 is equivalent to the Quadre de Classificació de l'Associació Catalana de les Universitats. Públiques (ACUP)
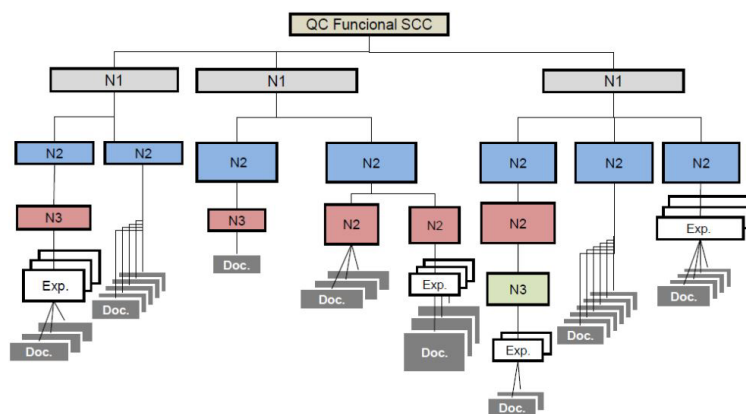

Figure 5: Multilevel representation of a QdC (Classification chart)

The drawing of the diagram responds to this evolution. But we have to see the evolution of the records keeping model according to the multientity model and then see the interoperability between systems.

For the identification of the academic record as a case study, it has been taken into account to model the diagram of the degree, the subject, the academic year, the group and the agents and devices.
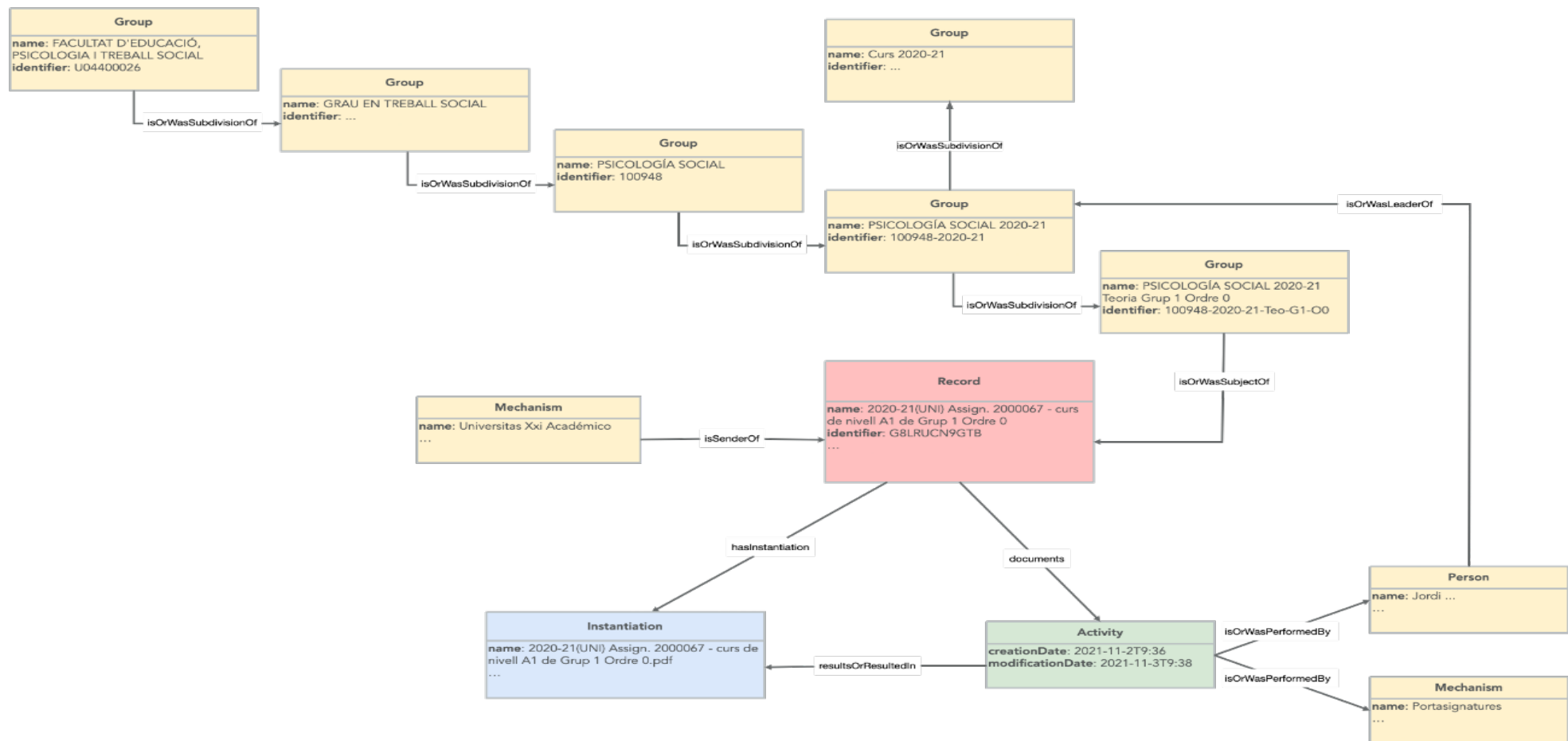
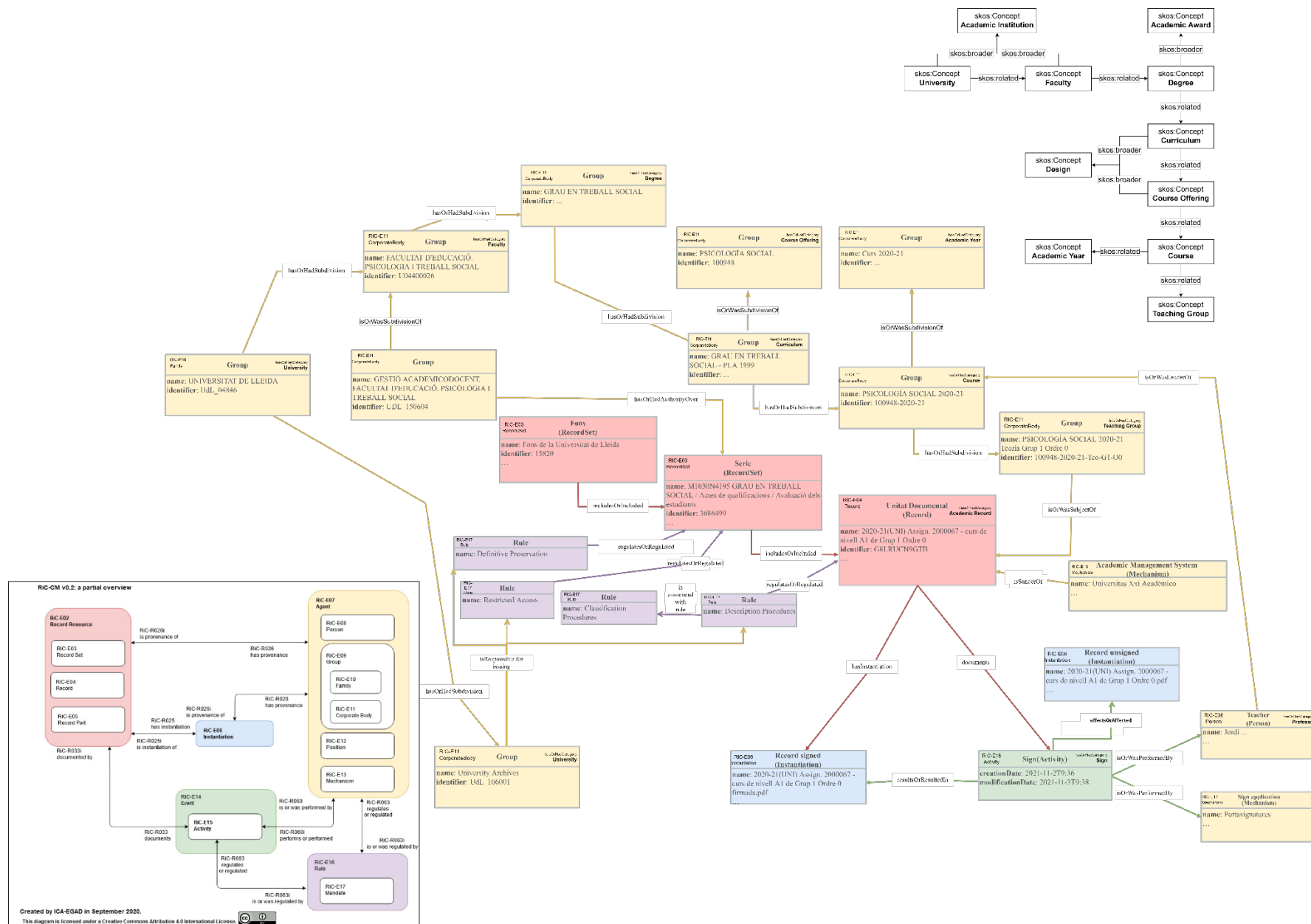**Diagram 1. The initial version of the adaptation of the RiC-CM to the academic proceedings**

Diagram 2. Diagram evolved according to Records in Context. Modelled according to the SKOS model (upper right) and the RiC model (bottom left)

The study focuses on the data for a single record. RecordSet are added for the found and series. Also, the agents for the UdL and academic management identified [indicated the Negotiated as a processor and the professor as author]. Regarding the rules, within the framework of the multientity model refer to the entity regulations. In addition to the access regulations, other regulations that affect the academic record must also be observed, such as new rules that regulate its content. Also very significant is the regulation of preservation, which due to the connotations of its content is permanently preserved and with restricted access. (D1. Restricted access, to contain mostly ordinary personal data. Personal data is not specially protected or merely identifying contained in public information directly related to the organization, operation or public activity of the Administration. This exclusion is void thirty years after the production of the record according to Catalan regulations).

The SKOS (Simple Knowledge Organization System) model and the RiC model are the models represented and only one sample for the entities that appear in the figure.
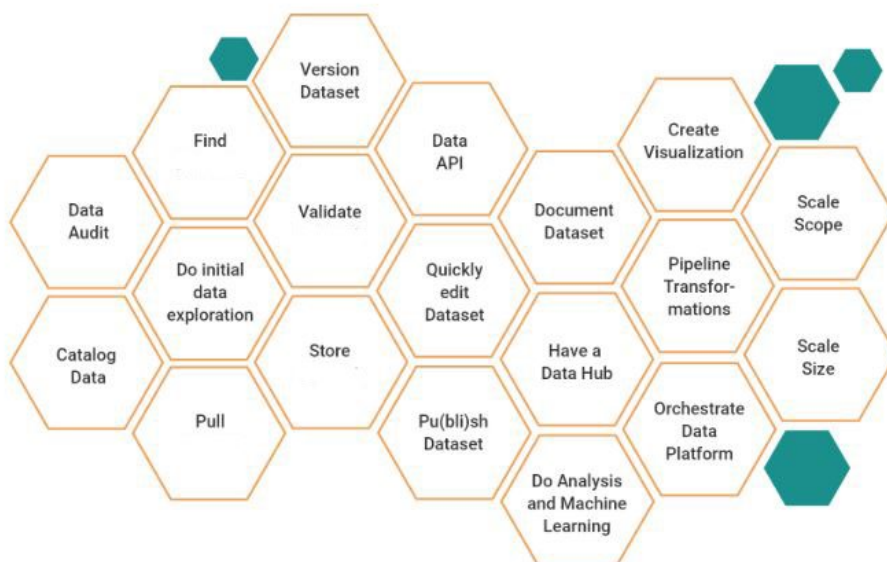
STUDY OF INTEROPERABILITY

The research is studying a solution on the interoperability of systems with the definition of a hexagonal architecture being promoted by Piql Iberia, and within a microservices architecture.

Microservices

Microservices are both an architecture and an approach to writing software. With microservices, applications are broken down into their smallest components, independent of each other. Instead of a traditional, monolithic, approach to apps, where everything is built into a single piece, microservices are all separated and work together to accomplish the same tasks. Each of these components, or processes, is a microservice. This approach to software development values granularity, being lightweight, and the ability to share similar processes across multiple apps.

This is at the interface of the ArkHive solution for digital preservation, a system based on a microservices architecture developed by Nageru Solutions, a subdivision of PIQL Iberia. Each hexagon represents an encapsulated functionality, that is, a microservice.

ArkHive has created an **abstraction layer,** a way of hiding details of a computing subsystem, allowing separation of concerns and interoperability. For example, the hexagon corresponding to the FIND is a connector that gets the files from any filesystem to be able to process them as a digital object .

3. CONCLUSIONS IN RESEARCH

The university's first need has been that academic records must be within the MMS to guarantee authenticity, integrity, reliability and veracity (ISO 15489 establishes it as a requirement), when they are records evidence generated from the academic manager and then in the signature application that incorporates this part of the evidence when the document is signed by the persons responsible.

These pieces of evidence need to be well informed within the EDRMS and to be preserved in the long-term archiving system is one of the most important objectives of the management system for documents according to ISO 30301 and also of this research where the need to build the interoperability of systems as a construction methodology in space and time was detected when these acts are generated and must be preserved with guarantees. The challenge in the study has been how it became possible. It has been demonstrated after a deep identification of the values of the metadata elements that the record when passing through diverse information systems, all different because their functions are different but must be interoperated with each other to establish a coherent construction in what has been defined in the title as an Archival Management System.

It has been observed that initially it has started from a multilevel description system that comes from a description style based on ISAD-G and that responds to a more hierarchical style of administrative organization and thus is also reflected in information systems, to face more and more a style of description that addresses diversification, and hence the need to explore a multientity representation according to ISO 23081 and following a graph model using Records in Context for a scenario specifically for the preservation of academic acts, with the added value of being able to capture the context and facilitate the future integration of different sources.

The next step of the research is to demonstrate through semantic technology how this graph representation allows to improve this interoperability and design a more realistic system to the needs of the organization based on this Archival Management System according to the renewal Recordkeeping Model where the standardization is allowing to approach to define more complex realities within the framework of Computing Archival Science.

BIBLIOGRAPHY

Metadata

- E-EMGDE (Esquema de Metadatos para la Gestión del Documento Electrónico)
  https://administracionelectronica.gob.es/ctt/eemgde
- PREMIS (Preservation Metadata Maintenance Activity) Library of Congress
  https://www.loc.gov/standards/premis/
- MEL (Metadata Extractor & Loader) Australian National University
  http://ceur-ws.org/Vol-2980/paper399.pdf
- ISO 23081:2021. Information and documentation — Metadata for managing records — Part 2: Conceptual and implementation issues. Geneve: ISO, p. 15

Semantic technology

- Semantic Web from w3 https://www.w3.org/standards/semanticweb/
- J2RM: an Ontology-based JSON-to-RDF Mapping Tool
  http://ceur-ws.org/Vol-2721/paper593.pdf
- AGRIF (The Australian Government Records Interoperability Framework) ontology
  http://linked.data.gov.au/def/agrif
- RiC (Records in Contexts). A Conceptual Model for Archival Description (2021). París: International Council on Archives.
  https://www.ica.org/sites/default/files/RiC-CM-0.1.pdf   [access 28/7/2022]
- SKOS (Simple Knowledge Organization System)
  https://www.w3.org/2004/02/skos/
- 2018 Semantizing and visualising archival metadata: the PIAAF French prototype online PIAAF (Pilote d'interopérabilité pour les Autorités Archivistiques françaises) https://www.ica.org/es/node/17606
  https://piaaf.demo.logilab.fr/

Interoperability

- Digital Preservation Archives – A New Future Architecture for Long-term Interoperability
  https://arc.aiaa.org/doi/pdf/10.2514/6.2018-2402
- Development of New Standards for OAIS Interoperability
  https://trs.jpl.nasa.gov/bitstream/handle/2014/52830/CL%2320-1268.pdf
- Data archive interoperability working group (MOIMS-DAI)
  https://cwe.ccsds.org/moims/default.aspx#_MOIMS-DAI
- 2017 Records in contexts: the road of archives to semantic interoperability
  https://www.emerald.com/insight/content/doi/10.1108/PROG-03-2017-0021/full/html
  https://core.ac.uk/download/pdf/132545189.pdf

Records in Context

- RiC (Records in Contexts). A Conceptual Model for Archival Description https://www.ica.org/en/records-in-contexts-conceptual-model
- 2017 Daniel Pitti - Records in Contexts (RiC) presentation
  http://www.alaarchivos.org/wp-content/uploads/2018/01/1.-Daniel-V.-Pitti.pdf
- Repositori de github amb alguns exemples de RiC-O de Matterhorn-Switzerland , NationalArchivesOfFrance i University-of-Strathclyde-Archives_Scotland :
  https://github.com/ICA-EGAD/RiC-O
- ICA -Section on University and Research Institution Archives - SUV
  https://www.ica.org/en/section-on-university-and-research-institution-archives-suv