# From the Secret Archive to open and fair access. Ways of modelling legal ecclesiastical data from the XVI and XVII centuries

**Albani, Benedetta**
albani@lhlt.mpg.de
Max Planck Institut for Legal History and Legal Theory, Deutschland

**Anokhina, Alexandra**
anokhina@lhlt.mpg.de
Max Planck Institut for Legal History and Legal Theory, Deutschland

**Park, Yohan**
park@lhlt.mpg.de
Max Planck Institut for Legal History and Legal Theory, Deutschland

## A complex starting point: from an inaccessible archive to FAIR data approach

The Vatican Archive is the private archive of the pope. Until 2019 it was called the Vatican Secret Archives and this particular name has given rise to legends, novels, films, rumours... Despite the fact that the archive has been open to the public since 1881 without any restrictions, and although the connotation of 'secret', which has always tickled the fancy of writers and journalists, derives from the ancient meaning of the adjective *secretum* , indicating that the holdings were the pope's private and personal property,[1] this archive and its very rich heritage remain inaccessible to many in several respects. This concerns both the archive holdings as a whole (archival fonds and series) and the historical documents preserved there.

In contrast to other historical archives whose inventories are based on International Standard Archival Description (International Council on Archives, 2000) that allow for interoperability of data, the Vatican Archive is so immense and complex[2] that it has not yet been equipped with a modern archival description system. Bear in mind that there is still no uniform archive guide, no comprehensive inventory, and that researchers often have to consult indexes and handwritten inventories, even dating back to the 14 [th] century, in order to know the contents of the archive series and select the documents they are interested in. Moreover, for reasons of document ownership and authenticity, the archive's policy is clearly resistant to an open access approach to data: documents cannot be photographed by researchers. Digital reproductions are expensive and can only be used for personal research purposes. Handwriting recognition software,[3] which usually foresee the sharing of digitized images with other users, cannot be used to 'read' these documents.[4] A final aspect that makes the documents preserved in the Vatican Archives hardly accessible concerns the specific knowledge and skills needed to read, understand and interpret them. Although common to historical research in general, this aspect takes on an important significance in the case of papal documents for the exegesis of which it is necessary to master certain specific disciplines and techniques.[5]

Among the various scientific objectives of our research project is also to improve the accessibility of data obtained from historical sources held in the Vatican Archive and to offer them to the scientific community according to FAIR principles (Findability, Accessibility, Interoperability, Reuse) ( Wilkinson *et al.* , 2016. ) , thus overcoming the inaccessibility of papal sources through modern technologies. In this paper we describe the methods and tools we are developing to address these challenges.

Our research project focuses on one of the most active bodies of the Roman Curia – the complex of organs and authorities that constitute the administrative apparatus of the Holy See - between the modern and contemporary ages: the Congregation of the Council, which we affectionately call "SCC"[6] . This dicastery was appointed for more than 350 years to oversee the correct interpretation and implementation of the Council of Trent[7] and the administration of justice around all disciplinary matters contained in the council and later on other papal laws. The Congregation of the Council was composed of cardinals and other personnel and met periodically to discuss and decide legal cases that came to it from all over the Catholic world. It had consultative, judicial and gracious functions and its jurisdiction ideally extended to the whole world. Based on the 1.5 km of the SCC archive, preserved in the Vatican Archive, the group has compiled a dataset describing approximately 35,000 *positiones* , that is judicial cases that took place in front of the SCC between 1564 and 1680 and involved thousands of people and institutions from all over the Catholic world and beyond.

In this paper we describe how we developed 1) modern standards for processing historical data, 2) the semantic model and 3) visualization strategies for contextualizing data in global history.

## Establishing modern standards for processing data extracted from historical sources

Working with early modern sources require a complex approach. Source criticism considers such factors as uncertainty, unclearness, gaps, damaged fragments, mistakes, ambiguities etc. as a natural part of the context, which needs to be considered in the analysis.

For various reasons, the digitization of the sources covered by the project was never considered either possible or desirable.[8] Instead, it was necessary to process the documents using specific techniques of source criticism applied to papal documents: all the data were collected directly from the original documents in accordance with academic standards of archival science, palaeography and papal diplomatics.[9] The group can today rely in an orderly and logically organized dataset composed by all of the approximately 35,000 *positiones* processed by the SCC from 1564 to 1680 and which constitute the oldest core of the dicastery's archive.[10] Alongside the data collection phase, the group systematized data and defined classes[11] in order to build the data model. For certain classes in the dataset the group also elaborated descriptive and structural metadata.[12] These processes have been carried out always making the interpretative intervention of the researchers recognizable. Thus, the dataset presents several 'double fields' of which one collects the data as it appears in the document and other the data in its standardized version. This, in order to allow the scientific community to criticize our interpretation and to enrich it with new elements in the future. These resources become the basis for creating the knowledge graph for our data and the visualization application described below.

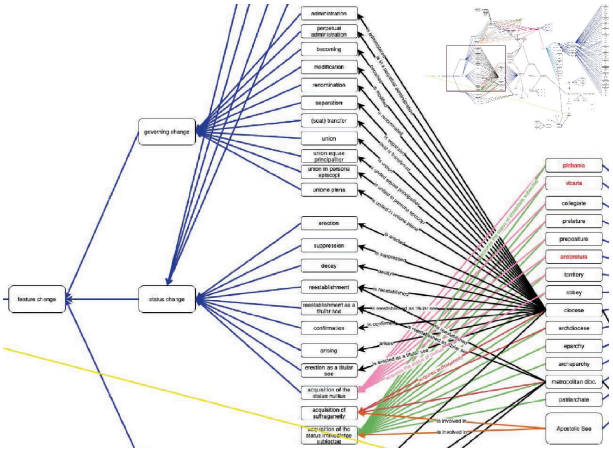


Figure 2: Section of the data model.

## Building the semantic model of the SCC administrative structures.

The research group currently curates and manages the data using a graph database for the purpose of knowledge management. We choose a NoSQL graph database, i.e. Neo4j, because of its flexibility, which allow to work with unstructured, semi-structured, and structured data , and create a schema-free model (Zhang, 2017). Figure 3 shows the data schema of *positiones* built in the Neo4j graph database. The individual nodes represent persons (entitled as Persona), places (entitled as Location), institutions (entitled as Dioceses), *positiones* (entitled as Positio). The archival sources are modelled with multiple nodes: *Volumen* (volume for *positiones* ), *Extra* (extra materials for cases), papal interventions (entitled as Intervention), papal fiat (entitled as *Fiat*), and *Nexus* (ID for connecting the different parts of a *positio* within the same volume and to indicate the links between different volumes. For each volume, a so-called *Vakat* node is created for collecting information about empty pages. The Node *Phase* is used to identify up to individual phases that a case may have. The edges represent the relationship structures between the nodes, such as the relations between a petitioner involved in a case and that case.

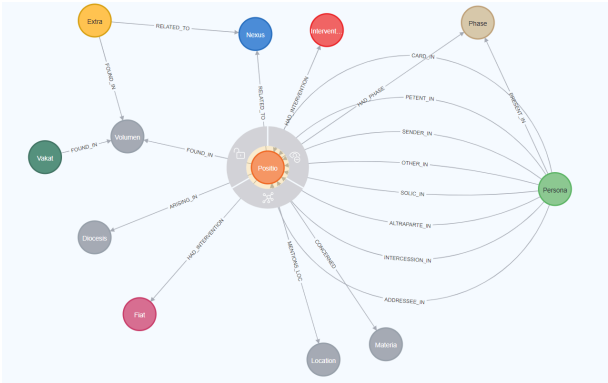


Figure 3: Data schema of *positiones* built in the Neo4j graph database.

## Modelling the legal administrative changes of the dioceses of the Catholic Church

In addition to usage of the graph database for curation of the *positiones* , the research group is in the process of developing an ontology for modelling the historical changes of the legal administrative institutions, which appear in the data, with special attention to dioceses, which are of fundamental importance in our context as they represent the basic object in legal administrative structure of the Catholic Church. These data will be available as a

knowledge graph. Using the Semantic Web technologies, we constructed an event-based ontology with key elements (Event, Time, Place, Institution, Source) that trace legal administrative events and changes of each diocese quoted in the *positiones* .

Our model proposes an extension of CIDOC-CRM (Doerr, 2003) through suggesting additional subclasses and object properties. Although CIDOC-CRM provides strong semantic expressiveness, this instrument was designed for modelling within the cultural heritage domain that set semantic limitations for modelling the structures of the administrative acts in historical perspective. Therefore, we proposed additional subclasses and object properties in order to fill a methodological gap.[13]

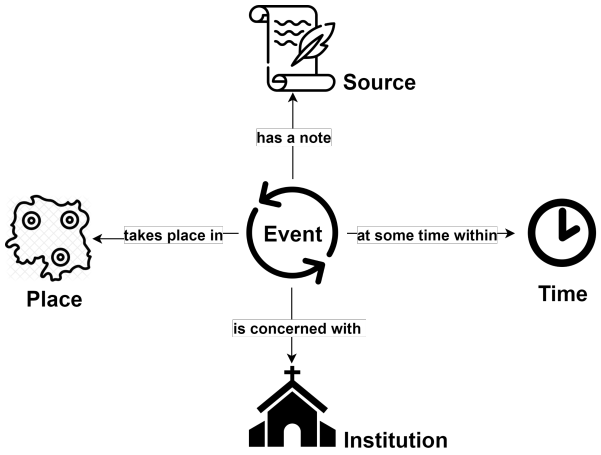


Figure 4: Key elements in our ontology

# Publishing open and FAIR data

Based on our ontology data model and in accordance with FAIR principles, we prepare the publication of our data as linked open data by using the combination of Resource Description Framework (RDF) triples and Internationalized Resource Identifiers (IRIs) which ensures for maintaining the semantic interoperability of data (Berners-Lee, 2009). Since for many of our actors, places, institutions there are no entries in any authority databases yet, therefore we will assign an URI to each resource and offer our data IDs as authority records for other projects.

In accordance with FAIR principles, this achieves and guarantees, on the one hand, the findability and accessibility of research data, and on the other hand, it enables researchers to use our data as a reference as well as for their own purposes. This provides also a fundamental framework for reusability of the data. We believe that providing our data in RDF triple format, we fulfil the requirements for interoperability and standardization of data according to the World Wide Web Consortium (W3C) recommendations. To achieve these goals, we currently identify and match entries from our dataset, su ch as dioceses and other actors, b y using the reconciliation web service API providing OpenRefine that enables to align datasets to entries from the already existing dataset[14] in OpenRefine ( Thalhath *et al.* , 2021). Therefore, we enrich semantically our dataset via authority data and achieve useful degree of interoperability and connectivity between our dataset and external data.

# Developing open source visualization instruments for contextualizing the SCC data and metadata in history

We consider that adequate historical contextualisation of data is necessary in order to avoid anachronistic or teleological biases and allow to interpret the results in scientific ways that correspond academic standards of historical research. To contextualize the SCC data and its metadata we developed visualization approaches that allow us to show our data not only as tables but also in interactive graphic format, to improve accessibility from the user perspective. There are various methods for visualizing historical data, focused on modern ideas of time[15] and space[16] that however don't count uncertainty, unclearness and incompleteness of historical data . We stress on methods of visualization of unclear, uncertain and incomplete data for reducing the impact of modern gaze on time and history. We also aim to remain the complexity of the sources from the perspective of an historian, i.e. preserve the original data as presented in the sources even if unclear, uncertain or incomplete. This part of our project aims to extend the functionality of this kind of instrument with setting up the custom controllers, impossible to maintain in already existed tools and therefore make uncertainty, unclearness and incompleteness visible and accessible as an important specifics of the data. The principal languages we use for this part of the project are R and JavaScript .

We developed the *SCC Timeline Explorer* web app in which the history of the SCC is placed in global historical context. Through the visualization of parallel timelines, this application allows to explore different aspects of the history of the SCC (evolution of competences, turnover of the personnel, frequency of the meetings) within global history (Global Legal History, History of the Roman Curia, Pontificates). We enriched the original data with descriptive metadata, which provide information in case of uncertainty, unclearness and incompleteness of data. Since many inputs have incomplete information, we decided to use vis.js[17] and R, for working with both – the data in an advanced way and the graphic visualization in dynamics, which also allows to visualize the uncertainty, unclearness and incompleteness in a functional (R) and aesthetic way (JS + CSS). We also combined the data and controllers, as we want to let users choose the settings.

Using R we created a reactive dataset, which connects the original dataset and controllers. Since the graphic part of vis.js works only with dd-mm-yyyy format, we added and external CSS file to stylize uncertain entities, which set the uncertainty as semi-transparent elements. For setting the controllers, firstly, in R we wrapped them as a reactive function. In a basic way, this solution uses shiny for calling reactive and dynamic functions in-

side a server part. Secondly, these functions are visualized in vis.js.

```
subset <- reactive({
  category <- paste0(c(input$sub_group_a,
                       input$sub_group_b),collapse = "|")
  category <- gsub(",","|",category)
  type_search <- paste0(c(input$type),collapse = "|")
  type_search <- gsub(",","|",type_search)
  dataset[grepl(category, sub_group) & grepl(type_search,type_of_act)]})

#setting the type controller
checkboxGroupInput("check_act_reform", "Type of the event",
choiceNames = mapply(type, icon, FUN = function(type, iconUrl),
SIMPLIFY = FALSE, USE.NAMES = FALSE),
choiceValues = c("act", "reform", "foundation"),
selected = c("act", "reform", "foundation"))
```

Figure 5: A basic example of setting a subset with controllers.


Figure 6: A timeline prototype using vis.js and R.

Each label has a trigger on click, which shows the information about the event or range, which is based on reactive values. What is unique in this approach is the possibility to set up reactive controllers for groups, custom reactive values, advanced aesthetics, and a way to bypass limits of the vis.js in the case of unclear, uncertain on incomplete information that is essential for historical datasets.

Products of our development will be published in the format of web applications in the SCC Explorer Platform. The commented code will be available on our GitHub. The results can be validated and repeated with other datasets. R functions are designed universal and scalable for other Digital Humanities requests.

# Conclusion

Our research project offers various methods of modelling and visualization of a large scaled database and metadata, considered in a FAIR way with open access, open data and open code. Firstly, we developed a knowledge graph using a semantic data model, which offers a data-centred approach for the Congregation of the Council in global context in the early modern period. Secondly, we created open access research tools for contextualizing historical data, including unclear, uncertain and incomplete information, into a big picture of global legal history, providing a graphic and accessible visualization of the data. For the time being, our data covers the period from 1564 to 1680 and are concentrated on the SCC, but data collection will continue for the later periods and our methods can be profitably applied also to other bodies of the Roman Curia. Therefore, our research and the tools we are developing can be of great importance for the understanding of the administration of justice in the Western World from the Late Middle Ages to Contemporary period.

# Fußnoten

1. On the history and heritage of the Vatican Apostolic Archive there is an extensive, though very fragmentary, bibliography. For an initial overview, see *Religiosa archivorum custodia. IV Centenario della Fondazione dell'Archivio Segreto Vaticano (1612-2012)* . 2015. Città del Vaticano: Archivio Segreto Vaticano and Gualdo, Germano. 1989. *Sussidi per la consultazione dell'Archivio Vaticano* . Città del Vaticano: Archivio Vaticano.
2. Today, the archive consists of more than 600 archival fonds from different types of institutions and covers approximately 85 linear kilometres of shelving.
3. For example the software *Transkribus* : https://readcoop.eu/transkribus/ .
4. The University of Roma Tre in collaboration with the Vatican Apostolic Archive in the frame of the project *In codice ratio* ( http://www.inf.uniroma3.it/db/icr/ ) is developing a software for text recognition of the volumes of the *Registra Vaticana* . This is an ambitious project, yielding excellent results, but unfortunately, due to the profound differences between medieval, early modern and contemporary writing systems, it will not be applicable to manuscript documents that are not written in the same style as the Vatican Registers. Firmani, Donatella, Paolo Merialdo, Elena Nieddu and Simone Scardapane. 2017. "In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks." In *11 th Italian Workshop on Artificial Intelligence for Cultural Heritage* ; Lastilla, Lorenzo, Serena Ammirati, Donatella Firmani, Nikos Komodakis, Paolo Merialdo, Simone Scardapane. 2022. "Self-supervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library." In *Information Processing and Management* 59(3).
5. For example papal diplomatics and specific branches of palaeography, sphragistics, chronology and chronography as well as the history of the Papacy and the Roman Curia.
6. In this paper, we will refer to the Congregation of the Council as 'SCC', an acronym used in specialist literature and derived from the Latin name of the institution: *Sacra Congregatio Concilii* .
7. The Council of Trent (1545-1563) was the 19th ecumenical council of the Catholic Church and remained in force until the 19th century. It was of central importance to the history of the Western world and beyond. Politically, the council attempted, unsuccessfully, to settle the rift between Catholics and Protestants that had arisen from Lutheran ideas by addressing important theological and ecclesiological issues (doctrine of justification, role of grace, existence of saints, doctrine of the sacraments, etc.). Within the Catholic world, the council constituted an important point of reference on a pastoral and juridical level. It is the council that remained in

force the longest of all the councils recognised by the Catholic Church (307 years) and thus left an important imprint on Catholic societies, an imprint that is still visible today.
8. This is due both to the very high costs of such an operation and other factors more related to source criticism such as the fragmentary nature of some of the cases decided by the SCC (some are divided into several phases also preserved in different volumes), the precarious state of preservation of some volumes, the extreme complexity of the structure of the processes (identification of the parties involved, the role of the cardinals who were members of the congregation, the institutions mentioned often related to canon law issues, the locations etc.) and the handwritings with which the documents were written (abbreviations, symbols etc.).
9. This phase of the work (2013-2019) was carried out by Dr. Benedetta Albani and Dr. Francesco Russo between 2013 and 2019 and was coordinated and financed by the Max Planck Research Group " Governance of the Universal Church after the Council of Trent " directed by Dr. Albani.
10. Vatican Apostolic Archive (AAV), *Congr. Concilio* , *Positiones* , 1-271.
11. Actors and their semantic roles (petitioners, members of the dicastery, lawyers and procurators, senders, addressees, sponsors etc.), institutions (dioceses, parishes, bodies of the Roman Curia, secular authorities etc.), places (spatial data, including coordinates, toponyms, etc.), temporal entities (events, dates, time periods), legal procedures, legal subject matters. For the moment, we have evidence of at least 8,000 petitioners, 1.500 places, 900 dioceses, 700 abbeys, 80 religious orders, 130 cardinals, 17 pontificates. These are preliminary results. Definitive data will be provided after the ongoing data cleaning process will be completed.
12. Biographical data of persons, metadata on the history of the mentioned institutions (dioceses, religious orders, churches, monasteries, abbeys), geographical coordinates of places and historical evolution of place names, bibliographical references, etc.
13. The limitation of the CIDOC model already shows in mapping on our data model, for example, the class crm: E8 Acquisition did not exactly fit our modelling notion, because the class was primary intended to design the legal process in the museum landscape, such as lending artwork to a gallery.
14. For example, some data were provided by the project *Monasteries, Collegiate churches, and Convents of the Holy Roman Empire and neighbouring countries* ( https://adw-goe.de/germania-sacra/klosterdatenbank/datenservice/ ), by the Göttingen Academy of Sciences.
15. For example the *Timeline JS* by Northwestern University Knight Lab ( http://timeline.knightlab.com ) or *MIT HyperStudio's Chronos Timeline* ( http://hyperstudio.mit.edu/software/chronos-timeline/ ) cannot visualize unclear and incomplete dates.
16. Almost all projects in geo-spatial visualization are based on GIS, which proposes a modern gaze on geography that does not allow to operate with unclear, uncertain and incomplete data in historical sense. For example, *Esri Story Maps* and the *Digital Humanities* projects ( https://collections.storymaps.esri.com/humanities/ ). On modern geography based visualization in R see Weinberg Eric. 2018. "Using Geospatial Data to Inform Historical Research in R." In *Programming Historian* 7.
17. The official website of vis.js library: https://visjs.org/ .

# Bibliographie

**Berners-Lee, Tim.** 2009. “Linked Data - Design Issues.” https://www.w3.org/DesignIssues/LinkedData.html

**Doerr, Martin.** 2003. “The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata.” In *AI Magazine* 24(3): 75. https://doi.org/10.1609/aimag.v24i3.1720.

**Firmani, Donatella, Paolo Merialdo, Elena Nieddu and Simone Scardapane.** 2017. “In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks.” In *11* th *Italian Workshop on Artificial Intelligence for Cultural Heritage*.

**Gualdo, Germano.** 1989. *Sussidi per la consultazione dell’Archivio Vaticano*. Città del Vaticano: Archivio Vaticano.

**International Council on Archives** . 2000. *ISAD(G): General International Standard Archival Description*. Ottawa: International Council on Archives.

**Lastilla, Lorenzo, Serena Ammirati, Donatella Firmani, Nikos Komodakis, Paolo Merialdo, Simone Scardapane.** 2022. “Self-supervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library.” In *Information Processing and Management* 59(3). https://doi.org/10.1016/j.ipm.2022.102875 .

*Religiosa archivorum custodia. IV Centenario della Fondazione dell’Archivio Segreto Vaticano (1612-2012)*. 2015. Città del Vaticano: Archivio Segreto Vaticano.

**Thalhath, Nishad, Nagamori, Mitsuharu, Sakaguchi, Tetsuo, and Sugimoto Shigeo.** 2021. “Wikidata Centric Vocabularies and URIs for Linking Data in Semantic Web Driven Digital Curation.” In *Metadata and Semantic Research*, ed. Emmanouel Garoufallou and María-Antonia Ovalle-Perandones, 336–344. Metadata and Semantic Research. MTSR 2020. Communications in Computer and Information Science, vol 1355. Springer, Cham. https://doi.org/10.1007/978-3-030-71903-6_31.

**Weinberg Eric.** 2018. "Using Geospatial Data to Inform Historical Research in R". In *Programming Historian* 7 . https://doi.org/10.46430/phen0075 .

**Wilkinson Mark D., Dumontier, Michel, Aalbersberg, I. Jsbrand Jan et al.** 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” In *Sci Data* 3. https://doi.org/10.1038/sdata.2016.18 .

**Zhang, Zuopeng Justin.** 2017. "Graph Databases for Knowledge Management." *IT Professional* 19(6): 26-32. https://doi.org/10.1109/MITP.2017.4241463.