

The Semantics of History. Interdisciplinary Categories and Methods for Digital Historical Research

Esther Travé Allepuz¹, Pablo del Fresno Bernal², Alfred Mauri Martí³, Sonia Medina Gordo¹ *

¹ Universitat de Barcelona, Barcelona (Spain)

² Sistemes de Gestió de Patrimoni SCCL, Barcelona (Spain)

³ Centre d'Estudis Martorellencs, Martorell (Spain)

Received 4 November 2020 | Accepted 21 December 2020 | Published 10 February 2021



ABSTRACT

This paper aims at introducing and discussing the data modelling and labelling methods for interdisciplinary and digital research in History developed and used by the authors. Our approach suggests the development of a conceptual framework for interdisciplinary research in history as a much-needed strategy to ensure that historians use all vestiges from the past regardless of their origin or support for the construction of historical discourse. By labelling Units of Topography and Actors in a wide range of historical sources and exploiting the obtained data, we use the Monastery of Sant Genís de Rocafort (Martorell, Spain) as a lab example of our method. This should lead researchers to the development of an integrated historical discourse maximizing the potential of interdisciplinary and fair research and minimizing the risks of bias.

KEYWORDS

Unit of Topography, Actor, Integrated History, Ontology, Database, Records Management, Landscape Archaeology.

DOI: 10.9781/ijimai.2021.02.002

I. INTRODUCTION

HISTORICAL Science is a discipline devoted to the analysis and comprehension of the past for a better understanding of the present, and hence a potential forecasting of the future. Therefore, space and time are unavoidably the main scenarios of research in History, and the alternation between permanence and change is its main object of study. This analysis of the so-called historical time [1], in which different entities exist and events occur transforming the reality of the past into something new and different, leads historians to construct the past in a narrative form.

The fact that data related to time and change are present in any written, material or immaterial vestige of the past boosts –and even forces– interdisciplinary research in History. Despite this, traditional approaches have frequently focused on written vestiges preferably, while disregarding other sources of information such as archaeology, iconography, literature, and a wide range of social sciences and humanities, which have been often considered as complementary at their best. In doing so, the different vestiges of the past do not *integrate* within a single discourse, but originate parallel discourses that might incur into contradiction and potentially lead to endless academic debates.

The spatial and material turns in History [2]–[6] have challenged these old-fashioned approaches, and have led to a more accurate construction of the past. A new interdisciplinary research framework

–in which the difference between SSH and STEM blur into a new paradigm of digital and FAIR science– challenges History, and requires an effort from different disciplines in order to explore common languages and codes for the construction of the past. In this process of science going digital, common and exchangeable units of information are required, despite the specificity of different areas of expertise. Within this domain, some experiences on NLP have been developed [7]–[8] and most of them are related to textual sources [9]–[10]. Despite this, the strengthening of concept-based and relation-based corpora for the development of NLP in digital humanities [11] focusing on the ontological approach to historical data suggests a challenging scenario for historians. Scholars dealing with historical science should face the need of rethinking methodologies and the way to use ICT in order to solve wider and more complex research questions and to take our investigation beyond the 20th Century historical issues. One of our objectives is to provide ontological reflections about data and data management in order to produce richer historical relates, as long as they integrate as many vestiges of the past as possible.

Our contribution aims at offering a methodological proposal and practical application our teamwork has developed within the last years, as an extended version of [12]. Arising from landscape archaeology and the study of the material vestiges of the medieval period, our research methodology deals with data labelling and records management, and nowadays it has overcome the archaeological domain to integrate all vestiges of the past regardless of their nature or origin to strengthen historical research in the digital domain. We will develop these issues within an updated state of the art in section II. In the following sections, we will describe the method according to the labelling categories we propose and the resulting data modelling. We will use a practical case of study –the medieval monastery of Sant Genís de Rocafort, Martorell, Spain– as an illustrative example.

* Corresponding author.

E-mail addresses: esther.trave@ub.edu (Esther Travé Allepuz), pdfsgp@gmail.com (Pablo del Fresno Bernal), bnn@heraclit.net (Alfred Mauri Martí), smedingo12@alumnes.ub.edu (Sonia Medina Gordo).

II. LANDSCAPE ARCHAEOLOGY AND INTEGRATED HISTORY

History has not been the most enthusiastic discipline to join the so-called Digital Humanities so far, and data managing strategies have been widely challenged in our domain [13]. Far from being overwhelmed by the unknowns of this digital turn, a few exceptions deal with different ways of representing historical information [14]–[15] and the building of a semantic definition for historical ontology [16]–[18]. Recent experiences focus on quantitative data analyses [19] and, predominantly, on written historical texts [20]–[21]; and some of them struggle to find the best ways to deal with bias [22] and uncertainty [23]. Despite this, a normalized user-friendly code to exploit vestiges of different nature and support is still missing and historical knowledge seems to be restricted to its written apparel.

Our team developed a first proposal to identify minimum units of information and label them as Units of Topography, Units of Stratigraphy and Actors, as will be defined below, in the framework of Landscape Archaeology. The initial goal was to integrate vestiges both from written or material sources in the archaeological analysis of medieval landscape and, therefore, to explore landscape as a historical construct from a holistic perspective. More than twenty years later, our research information system has gone far beyond landscape archaeology or the medieval period to become a solid proposal for historical research, understanding Historical Science as a FAIR [24] construction of the past. It follows the principles stated by far-reaching research projects in this field such as the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) [25] (pp. 6–7) or Parthenos [26]. This FAIR-ness is achieved when ensuring the reliability and traceability of the research process, and when integrating as many sources of information as possible, even those that had been traditionally disregarded until recently as marginal or non-significant.

This is possible due to the development of ICT in the past decades. Recent advances allow for a significant speedup of data gathering and exploitation processes of much larger datasets, which opens a brand-new field for historical research in which new and more complex questions can be asked to past vestiges. Ensuring the existence of clear and unambiguous definitions of the ontology-mediated elements that identify units of information and their relations [27] is an underpinning issue to this regard. The following section deals with it.

III. LABELLING CATEGORIES AND DATA MODELLING

As defined by K. Thibodeau [28] (p. 7), an Entity is something that existed and an Event is something that happened or was done. Entities

and Events have a relationship of involvement, as every event involves at least one entity that might be the participant in the event, its observer, the mechanism for the event to happen, or the object altered by the event itself. In terms of data-labelling, the categories Unit of Topography and Actor, as defined by A. Mauri [29] (p. 45), and their relations, provide the unique and univocal identifiers for historical facts regardless of their link to permanence (Entity) or change (Event), or the nature and support of the vestige. Units of Topography as we use them are, in fact, a wider conception of archaeological Units of Stratigraphy [30], which overcomes their materiality and turns them into a broader concept to identify any entity or action existing or occurring at a particular time, notwithstanding its presence or absence in the archaeological record.

The following definitions apply to each one of these categories:

- Unit of Topography (UT): It is the evidence of an action or situation that can be located in space and time, regardless of the specificity of the information source and its biotic, non-biotic or anthropic attributes. Each UT has a specific location and date. Location can be expressed as a UTM coordinate or as an administrative delimitation that might have changed through time.
- Unit of Stratigraphy (US): It is the material evidence of an action occurred in the past, representing an archaeological aspect of the cycle of time. They are of universal character and can be found on any archaeological site in the world [30] (p. 42). As a reflection of materiality, graphic and cartographic representations are essential attributes of these units.
- Actor: It is the individual or corporative, active or passive, protagonist of an action identified as a UT. If being an individual, its attributes are their name, gender, religion, citizenship, date of birth and death, etc. Different individual actors gathered for a given period of time with a particular purpose and under determinate conditions can act as corporative actors.

As the US category is contained in the definition of a UT and we might consider them as equals at some point, Table I summarizes their differences and ontological specificities.

As shown in Fig. 1, several types of relationships can be set between UT/US and Ac. A UT can include, link or delimitate another UT. Hence, Inclusion, Delimitation and Link are classes of the UT-UT relation. An Actor always plays an active or passive role within a UT, so Role is the only class of Ac-UT. Actors can relate to other actors through familial, political, social or economic Ac-Ac relationships. The materiality of US implies that the only possible relation between US is physical contact. When interpreting the archaeological register, we can group several US into activities and assemble these activities into groups of activities

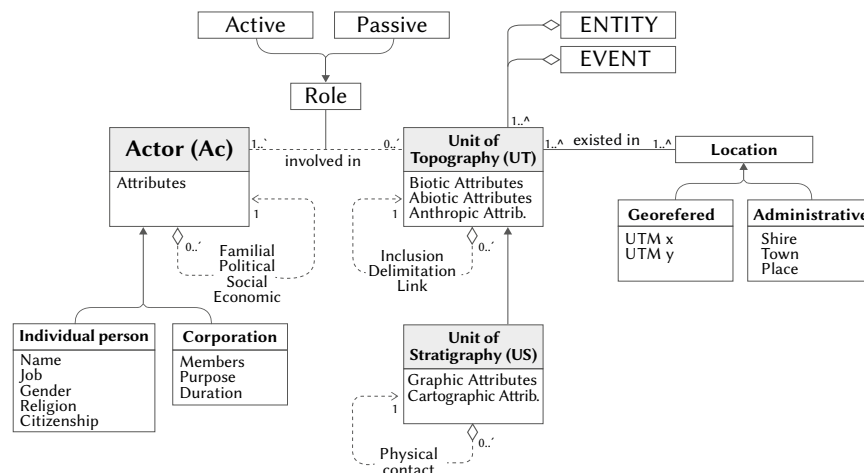


Fig. 1. UML diagram of ontological concepts –UT, US, Ac and their relations– for integrated historical research.

according to [31]. The relation and equivalence between UT, as shown in Table I, and US is then a matter of scale: UT can be equivalent to US in the archaeological record, or we can define UT during the data interpretation process. Anyhow, the UT scale is the one allowing for an interdisciplinary approach in History.

TABLE I. UNITS OF INFORMATION (UT/US) ONTOLOGICAL SUMMARY

Attribute	US	UT
Source of obtention	Material	Written, material, visual, sound...
Materiality	Essential	Non-essential
Informative dimensions	Spatial	Essential
	Descriptive	Essential
	Graphic	Essential
	Cartographic	Essential
Temporal	Essential	Essential
Possible relations	US-US/UT-UT	Physical contact
	Ac-US/Ac-UT	Non-possible*
	US-UT	US = UT US + US = UT
Involvement in event	Essential (altered)	Non-essential

* The material register does not inform about Ac.

IV. A CASE STUDY: RESULTS AT SANT GENÍS DE ROCAFORT

A. Site and Materials

We have selected a Catalanian monument –the monastery of Sant Genís de Rocafort– located at the town of Martorell (Barcelona) in order to put into practice our labelling proposal and demonstrate the validity of our information system from an interdisciplinary perspective. The Lord of Castellvell and his wife founded the monastery of Sant Genís in 1042 as the ruling centre of a small Priory [32].

Since then, the monastery, located on the top of a hill, became an emblematic element of Martorell’s landscape, and down on the foothill the ancient temple of Santa Margarida became the parish church of the Priory, which had a flourishing period in 11th – 13th Cent AD, even though the monastic community never included more than four or five monks. The monastery became dependent of larger monasteries within a Benedictine network, probably in the 13th Cent. Since then, the monastery of Sant Miquel de Cruïlles (Girona, Spain) [33] had the patronage over Sant Genís, while being itself dependent from the Piedmontese Saint Michael’s Abbey (San Michele della Chiusa, Italy). In the late middle age, Sant Genís started its slow decadence until 1534, when it became a secular Priory owned by a Barcelonense merchant named Joan Bolet. He slightly refurbished the buildings of Sant Genís and Santa Margarida, and the Priory endured until the 19th Cent AD. After some years in private hands, the Priory became a public property of Martorell’s Town Council in 1967.

Historical research at this site has benefited from the proposed system of management information. Architectural analysis of the building, restoration works, and archaeological fieldwork provide a new research perspective for the construction of an integrated historical discourse built from the written evidences kept in archival records and thoroughly analysed in the past [34].

B. Practical Examples of Data Labelling

Accordingly, we selected some written, graphic or material vestiges of Sant Genís’ past and analysed them by identifying and labelling the US, UT and Ac informed within at different stages. Managing information throughout these lab examples in an integrated form

provides a synthesis for the method’s development. The following subsections deal with each type of selected sources, labelling them appropriately, and Table II and Table III summarize the information gathered so that it can be furtherly discussed.

TABLE II. SIMPLIFIED AC DATASET GATHERED FROM SELECTED EXAMPLES.

Ac	Name	Attributes	Related UT/Ac
01	Bonus	Lord of Castellvell	Ac02, 03; UT01
02	Guilielmus		Ac01
03	Sicardis		Ac01; UT01
04	Clement VII	Pope	UT20, 21
05	Joan Bolet	Barcelonense merchant	Ac06; UT07, 22, 23
06	Simó Capellades	Priest, Prior	Ac05; UT23, 25

TABLE III. SIMPLIFIED UT DATASET GATHERED FROM SELECTED EXAMPLES

ut	Brief Description	Related UT/US/Ac	Attributes	Date
01	Donation	Ac01, 03	Event (property)	1042
02	Romanesque Church	UT07, 25, 26, 27, 28, 29, 30, 31*; US5001 – 5024	Building	12 th Cent
03	County of Barcelona	UT07	Political entity	
04	Castellvell	Ac01; UT05, 06, 07	Town	
05	Martorell	UT04, 06, 07	Town	
06	Priory border	UT07, 08 – 18	Border	
07	Priory of Sant Genís	UT04, 05, 06, 22, 25	Religious entity	
08	Congostell	UT07	Place	
09	Mountain range	UT07, 10, 11	Mountain range	
10	Rosanes	UT07, 11	Place	
11	Rocafort	UT02, 07, 10, 25	Hill	
12	Montgoi	UT07	Hill	
13	Vena	UT07	Place	
14	Grau	UT07	Place	
15	Torrent	UT07	Place	
16	Lloreda	UT07	Place	
17	Torrent of Lloreda	UT07	Waterflow	
18	Anoia River	UT07	Waterflow	
19	Donation letter	Ac01, 03; UT01	Document	1042
20	Papal bull	Ac04, UT21	Document	1534
21	Concession	Ac04, UT20, 22	Event (gift)	1534
22	Patronage	Ac05, UT07, 22	Entity (property)	1534
23	Refurbishment	Ac05, UT02, 25, 31*; US247, 248, 249, 5025, 5027, 5031, 5033, 5068 – 5070	Building transformation	16 th Cent
24	St Miquel de Cruïlles	UT25	Religious entity	
25	Monastery of St Genís	UT07, 24, 25, 31*; US5026, 5030, 5032	Religious entity	
26	Partial collapse of Romanesque building	UT02; =US5034	Destruction	1448
27	Apse walls	UT02, 26, 28		
28	Apse demolition	UT02, 27; =US5086	Destruction	1928
29	Abandonment	UT02, 07, 25; US5019, 5023, 5077, 5080, 5083	Enduring event	19 th Cent
30	Restoration	UT02; US5035 – 5039, 5079, 5082, 5085	Building transformation	2014
31	Archaeological fieldwork	UT02, 25, 26, 28, 29; US247, 248, 249	Event (Research)	Since 2010

* Units of Stratigraphy (US) with the labelling format **000** are related to the Romanesque church or the monastery buildings and thus included in UT30.

1. Archival Sources or Written Primary Files

A copy of the donation document of 1042 kept at the Diocesan Archive of Barcelona [32] and published in [34] (p. 139-143) informs about the location and borders of the Priory. We can label texts and identify the units of information contained and their attributes by following this code: <UT00> <Ac00> <Att-UT00> <Att-Ac00> <Date-UT00>. We also label relations in accordance with concepts related: <Ac00-UT00> <Ac00-Ac00> <UT00-UT00>. A semicolon separates different UT, Ac, Attributes, or Relations identified through the same word or syntagmatic expression.

Sit omnibus notum quod ego <Ac01 Bonus>, <Ac01-Ac02 filius> <Ac02 Guilielmi>, <Att-Ac01 dominus Castri Vetuli>, et <Ac01-Ac03 uxor mea> <Ac03 Sicardis> pariter in unum in nomine Domini <UT01 donatores> Domino Deo et sancto Genesio martiri Christi, cuius <UT02 ecclesia> sita est in <UT03 comitatu barchinonensi>, <UT02-UT04; UT02-UT05 intra terminos> <UT04 Castri Vetuli> <UT04-UT05 de> <UT05 Martorello>. [...] <UT06-UT07 Habent autem> <UT06 terminum> <UT07 hec omnia> ab ortu solis in ipso <UT08 Coangustello>, atque ascendendo in sumitate <UT09 serre> pergit per sumitatem illius <UT09 serre>, que est <Att-UT09 intra> <UT10 Rodanes> <UT10-UT11 et> <UT22 Rocam fortem>, usque in <UT12 muntem Gaudii>. Et inde pergit usque ad <UT13 Bennam>, secundum quod hiemali tempore sive pluvioli decurrunt aque ab oriente contra septentrionem. A meridie in ipsa Bennam, atque pergit inde usque ad ipsum <UT14 Gradum>, et descendit in ipsum <UT15 torrentem> qui discurrit. Ab occidio in ipsa <UT16 Laureta>, et inde pergit per ipsam <UT17 rieram de Laureta> usque in medium <UT18 flumen Anolle>. A circio similiter in medium <UT18 flumen Anolle> et inde pergit per medium ipsius fluminis usque in ipsum <UT08 Coangustellum>. [...] <Date-UT01; Date-UT19 Facta> <UT19 carta donationis> huius sex idus aprilis anno undecimo regni Henrici Regis. [26], [29] (p. 140)

2. Photographic Vestiges and Architectural Analysis

Architectural analysis from an archaeological perspective aims at identifying the building phases and further transformations in architectural heritage. Photographic vestiges usually allow for the identification of constructive elements that have disappeared nowadays. We identified and labelled US in Fig. 2 accordingly:

3. Archaeological Fieldwork

Because archaeology is a destructive process, building a precise and detailed archaeological record following a clearly stated protocol has been the commitment of archaeological science for a very long time. US have been the main unit of information since 1980 [30].

In Fig. 3, we selected just three US for labelling as a lab example amongst the entire archaeological record at this site. The stratigraphic method for archaeological excavation and register is widely accepted among scholars in this domain and our UT/Ac labelling strategy is built in accordance to the archaeological method, as stated above, due to the authors' archaeological background. Therefore, we could not avoid selecting archaeological examples for building an integrated historical narrative, as the archaeological record is concomitant with our proposal for information management, even though the excavation results in Sant Genís are much wider than shown in this paper.

4. Bibliographic Reflections

In all scientific production and in any form of Past Construction in particular, the so-called state-of-the-art –or past reflections, in Thibodeau's terms [28]– are valuable sources of information that must be considered in terms of data labelling and management. The same labelling method proposed for archival sources or written primary files works for secondary information as well.

In Fig. 4 we show an example of data labelling within an excerpt from a published piece of research about the Priory of Sant Genís [34]. That is the summary and study of a rich documentary assemblage informing about the priory from archival sources, and providing the historical framework and state-of-the-art before the archaeological excavation started.

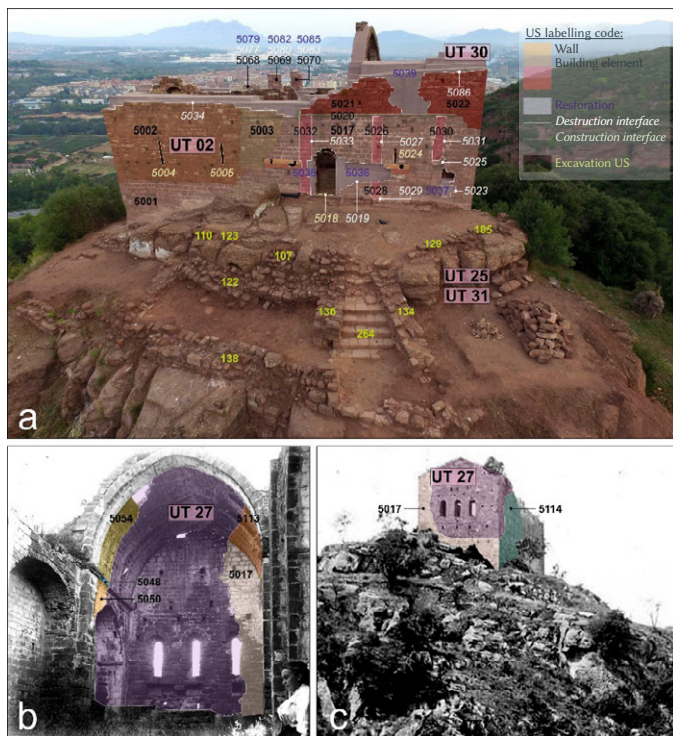


Fig. 2. Aerial view of Sant Genís de Rocafort after fieldwork carried out in 2020 (a), and photographic vestiges from the beginning of the 20th Century (b-c) [35]–[36]. Interpreted relations between US and UT are recorded in Table III.

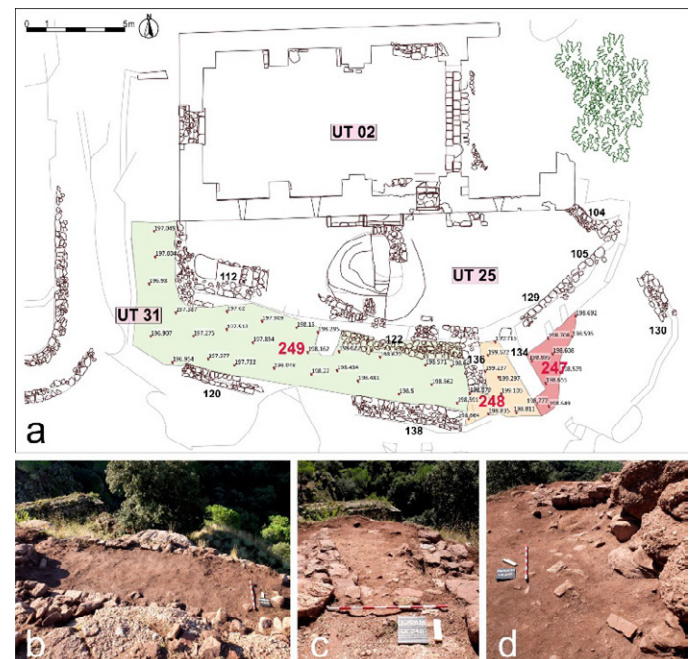


Fig. 3. Cartographic (a) and photographic representations of US 249, 248 and 247 (b-d). As per architectural analysis, US-UT relations are recorded in Table III.

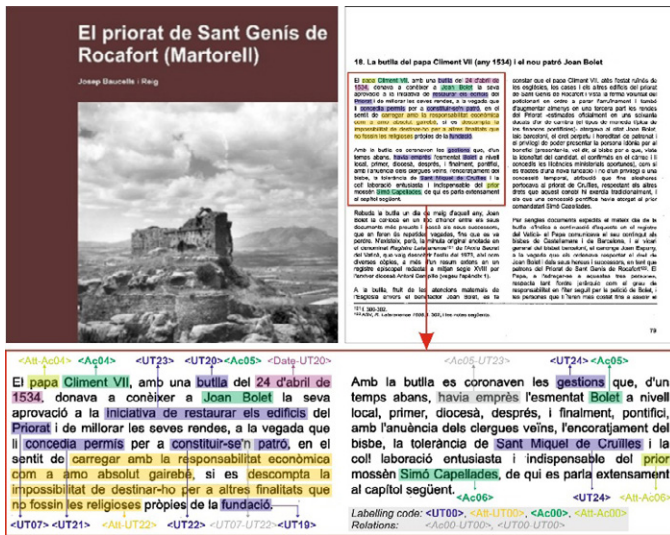


Fig. 4. Labelling example of a short excerpt from the bibliographic reference of a previously published study [34].

UT/Ac data labelling is operational both on original texts when possible –even if annotated by hand, as in the example on Fig. 4– and also on transcriptions and translations. In section I above, we labelled on the transcribed Latin version of the Priory’s foundation and below, the translation to the excerpt in Fig. 4 is labelled in the English language. Actors labelled in both examples are summarized in Table II.

In <Date-UT20 1534, April 24th>, <Att-Ac04 Pope> <Ac04 Clement VII> announced in his <UT20 papal bull> to <Ac05 Joan Bolet> his approval of the initiative to restore the buildings of the Priory and to increase its rents. He also <UT21 allowed> Bolet to exercise his <UT22 patronage> <UT22-UT07 upon the> <UT07 Priory>, <Att-UT22 assuming all financial responsibility as an owner>. The sole <Att-UT22 exception to this absolute control was to use the Priory for non-religious purposes>, as stated in the <UT19 donation text of 1042>, which was strictly forbidden. Receiving this papal bull is an achievement for <Ac05 Joan Bolet>, <Ac05-UT23 who had begun> the <UT23 arrangement of refurbishing work> at the <UT07 Priory> buildings some time earlier, with the collaboration of neighbouring clerks and bishop’s encouragement. The Benedictine abbot of <UT24 Sant Miquel de Cruilles> was tolerant with Bolet’s plans, and the <Att-Ac00 <Ac06-UT25 prior of> <UT25 Sant Genís>> <Ac06 Simó Capellades> was an enthusiastic and indispensable <Ac06-UT23 collaborator> <Ac06-UT05 of the> <Att-Ac05 Barcelonese merchant> as well. [34] (p. 79)

5. Landscape Analysis and Cartographic Sources

The proposal introduced in this paper offers a useful tool for landscape archaeology as well, and we can label cartographic sources similarly. We must bear in mind that the concept of Unit of Topography –in accordance with the definition proposed– implies a location and date as main attributes for further exploitation. UT/US gathered from selected examples and included in Table III have a precise location expressed as UTM coordinates that have been used to produce and label the maps shown in Fig. 5.

To that extent, Geographic Information Systems (GIS) have a great potential for data processing and exploitation, and they provide a useful tool for the landscape approach to historical knowledge. Providing topographic and chronologic attributes for past entities and events is a conceptual requirement for a spatial turn [37] in History, but also for a general scientific procedure of Past Construction under normalized terms and categories.



Fig. 5. Aerial view [38] of the Priory of Sant Genís with UT labelling according to the toponymy mentioned in the foundation document (a). Below, UT identification on a general map [29] of the Medieval County of Barcelona (b).

C. Data Exploitation

Cartographic representations and data exploitation by means of GIS technology are both a way to process data and a final representation which may be used for publication and dissemination. Anyway, our proposal –originally thought from the domain of landscape archaeology– goes beyond spatial representations and can focus on temporal sequences or relational data interpretation as well.

Although we summarized data gathered from our lab example in a couple of simple tables, these form part of an ontology-mediated database in which UT, Ac and the relations between them are collected in separate tables. The diagram shown in Fig. 6 represents the main components of a database storing the crucial units of information, their attributes and relations. When keeping this structure, databases can adapt to the needs of a particular research project and show variable interfaces and self-search exploitations, but they will always be interchangeable and potentially interconnected, as far as they share a common data modelling [39]–[40], as shown in Fig. 6.

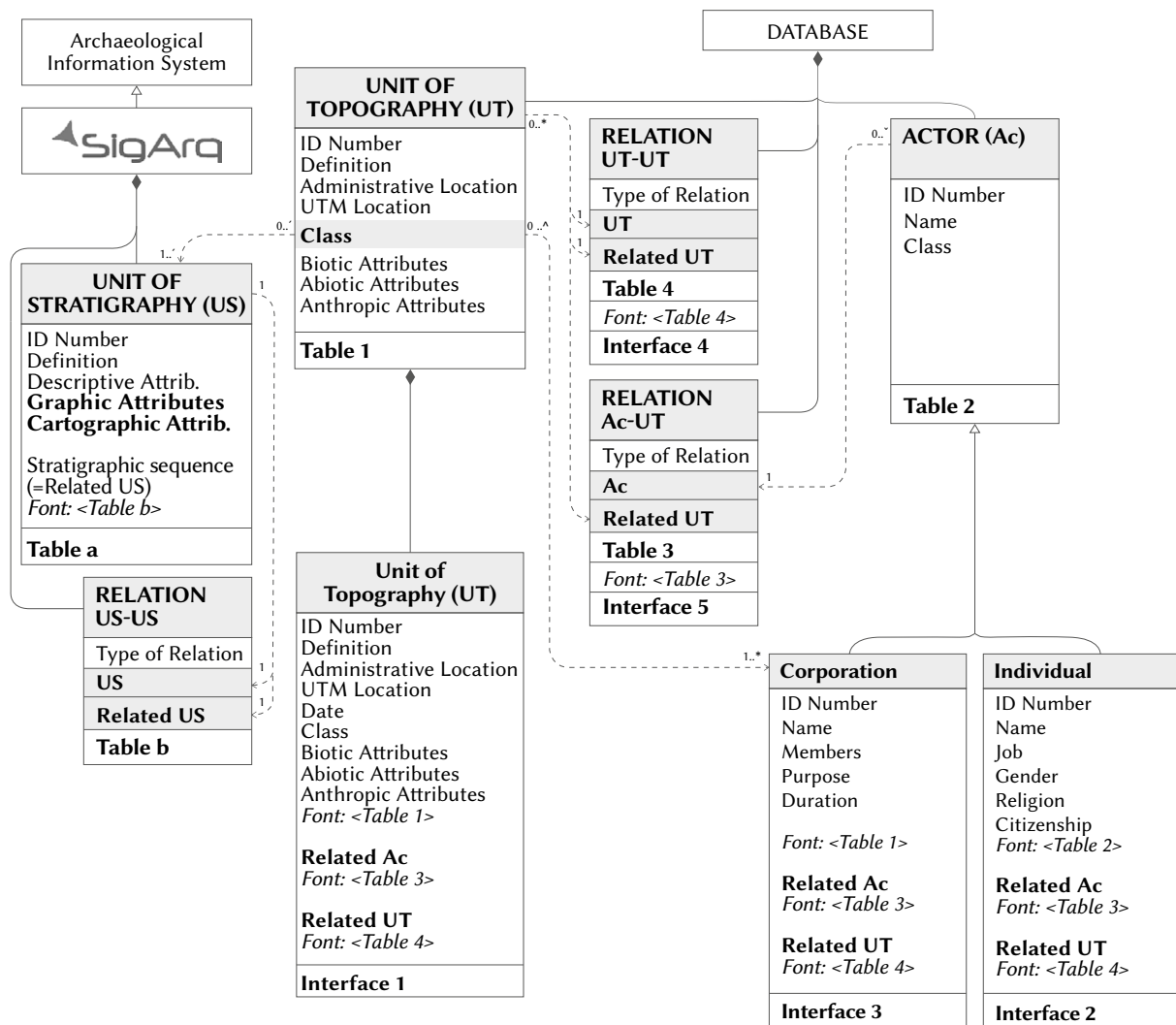


Fig. 6. UML diagram of generic database components.

Fig. 7 represents some interfaces of the current applications used nowadays by our team. One of these is a GIS-based information system created for archaeological purposes [41]–[42]. All of them share the same categories –UT/US/Ac– and hence they allow interdisciplinary research beyond the particular expertise of anyone of us.

When gathering and storing data in the form of tidy-structured tables with variables in columns and observations in rows [43], and according to identified US, UT and Ac regardless of the nature or support of the vestige, multiple representations are possible. Flux diagrams and matrices can visually establish the temporal sequence of activities and their permanence or transformation. Fig. 8 shows an extract of a historical Harris-like [30] matrix created for the Priory of Sant Genís the Rocafort including the Ac, UT exemplified throughout this paper. The archaeological US Harris-matrix routinely developed in archaeology is also included.

This is an example of data exploitation and representation as an interdisciplinary historical matrix, including the archaeological results within the historical discourse arising from written evidence and explaining the Priory’s past in a richer construction. Notice how the material vestiges of the buildings’ refurbishment in the 16th Century were positively identified within the archaeological register combined with the architectural analysis. In such a representation, Actors mentioned in written vestiges can be assigned to phases and located visually within the corresponding period.

V. DISCUSSION: TOWARDS AN INTEGRATED HISTORY

The most striking point of using Unit of Topography and Actor as ontological concepts of Historical semantics is that they allow for a truly interdisciplinary research. Unfortunately, today historical science understood as a whole still lacks a common code for data integration within its discourse. The methodological particularities of each method make sense as far as they follow specific goals and socially determined functions. This should not be a problem for creating an integrated construction of the past, as far as they share a common system for information management and exchange, which –unfortunately– has not happened yet [44] (p. 41-42).

In a context of FAIR research, the aim to create an integrated historical discourse is a challenge that historians should face with a sense of urgency. Nevertheless, interdisciplinarity in history does not mean –or should not mean– juxtaposing different past constructions arising from each discipline (history, archaeology, literature, iconography, archival science, linguistics, law, and SSH in general), but creating an interdisciplinary narrative joining the efforts of many different scholars. Sharing a common system for information management and exchange allows us to monitor the research process from the beginning and to locate information precisely, which makes it findable and accessible to colleagues from diverse expertise domains. This is the best way to deal with bias and uncertainty.

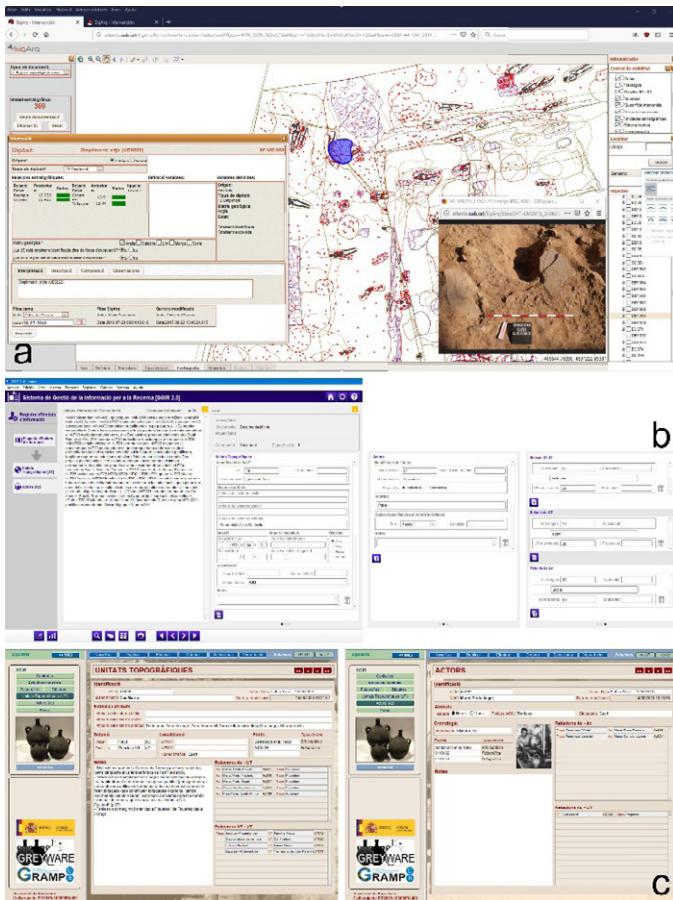


Fig. 7. Examples of record management tools developed by our team. Interface of SigArq software (a); SGIR 2.0 database screen view with forms for UT (left), Ac (centre) and Relation -UT-UT, Ac-UT and AC-AC- (right) (b), and Greyware database interfaces for UT (bottom left) and Ac (bottom right), showing in both cases relations between them as an automatized search from relation tables (c).

Our proposal suggests widening the scope of the archaeological method as a response to this challenge. As defined in [30] (p. 42), the concept of US includes any kind of action leaving a material imprint and identified within spatiotemporal coordinates, no matter if it is positive –adding materials– or negative –removing them. Therefore, during the archaeological fieldwork we register positive US when stratigraphic accumulation of materials occurs and we identify negative US in holes, broken structures or eroded layers.

This concept has proved to be wide enough to be adopted by other archaeology-related disciplines such as architectural analyses of buildings and material heritage studies [45] (p. 79). Since materiality – and, therefore, its cartographic and graphic informative dimension– is the main feature of US, could we define a similar concept equivalent to this unit of information but delinked from its material component? Yes, we could. Units of topography provide this univocal identification of entities and events in the past, with spatiotemporal coordinates and relations between them. Archaeology does not inform about actors, but many other sources of historical information do. Therefore, the proposal of UT/Ac gathering is an adequate compromise solution in order to develop an ontology for past construction in which entities and events are identified through non-ambiguous parameters.

Successful data labelling strategies (TEI) are limited to written sources and hence increasingly used in literature and language studies [46]. They might be successfully applied to the written examples we provided, but they fail in labelling iconographic or photographic vestiges. Textual encoding and labelling tools have a great potential

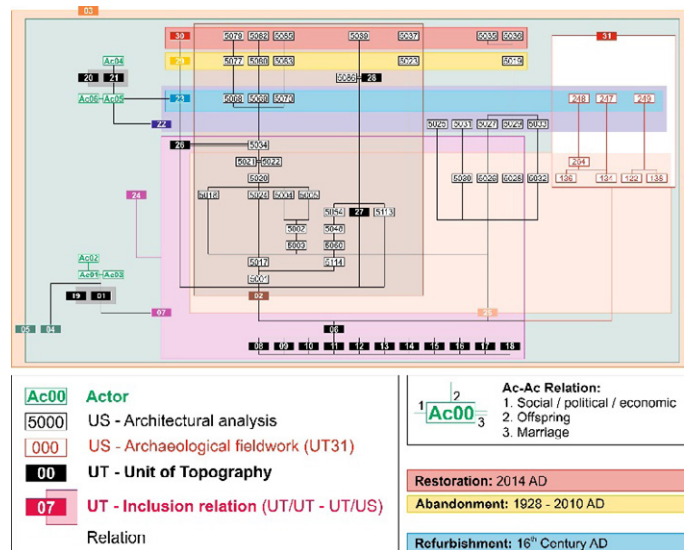


Fig. 8. Excerpt of the Historical matrix created for the archaeological site of Sant Genís de Rocafort. Only UT, Ac and US mentioned throughout this paper as a lab example for the method proposed and summarized in Table II and Table III are included.

for written vestiges from the past, but historical science needs an additional category for data labelling regardless of the origin or support of the vestige. Our labelling proposal implies looking for UT and Ac in too many different shapes and supports –even if textual sources are the most abundant. This actually implies more interpretative knowledge on the historians’ part, as it is not possible to detect these data units through mere automatic data labelling applications yet.

Furthermore, UT/Ac identification allows for multiple readings of past vestiges, which can be as exhaustive as required in a particular research project. Data gathered in the examples provided throughout this paper have been enough to demonstrate the validity of the method. We have attempted to find a balance between a theoretical demonstration and a practical case of study, but the historical sources informing about the Priory of Sant Genís de Rocafort are much wider and the historical matrix arising from them is far more complex. This also shows how the historian can read, analyse and interpret past vestiges to a desired level, according to their interest or domain of expertise, and how future experts dealing with the same vestiges can then generate new knowledge *building upon* previous reflections, but not *disregarding* them.

In recent years, data modelling and database construction in the terms described in this paper have allowed us to develop integrated approaches [29], [47] and software [42] overcoming the traditional inconveniences arising from the fragmentation of sources of information. Interactive multimedia and artificial intelligence have a great potential to automatize research processes and have proved to be novel and useful in the domain of SSH. Research projects in the field of History, Archaeology or Archival Science can benefit significantly from shared and transdisciplinary approaches to the past when using a common code. We consider the dialectics between US/UT and Ac as useful categories for data modelling, according to the semantics of Entity and Event as major ontological concepts in historical science [28], [48].

VI. CONCLUSION

Historical science is a wide discipline that has to consider all the sources of information available, which implies several other disciplines taking part in this process. Archaeology, linguistics, literature, and many others provide valuable data to contribute significantly to the

construction of the past. Historians should not add some of these data to a main discourse arising only from written sources but integrate all this information within interdisciplinary processes of data gathering and exploitation.

The NLP community has presented several attempts to process historical knowledge, according to two underpinning –one theoretical and one practical/methodological– ideas: the definition of event and data extraction through text labelling. In doing so initiatives share the common limitation of not having a precise and shared definition of event amongst the academy, and the data extraction procedure being limited to written sources. Furthermore, no attempt has been made to find a domain-specific definition of event combining the historical perspective and ongoing research in the NLP field.

Hybrid intelligence would be, to our perception, a challenging field to explore the possibilities of historical knowledge to become digital and interdisciplinary, and to develop appropriate UT/Ac recognition patterns. NLP systems might be focussed on finding and tagging event-meaningful concepts in written sources, even including archaeological excavation documents and text-supported records, but the heterogeneity of supports and formats for historical vestiges are much wider than these.

The concept of Unit of Stratigraphy, broadly used in archaeology, provides a useful characterization of actions in the past according to their materiality. Widening this idea, the categories Unit of Topography and Actor, as described throughout this paper, provide single and univocal semantic concepts to identify entities and events. Building databases according to these categories is a valuable strategy that integrates knowledge both from SSH and STEM to the historical domain, and made information systems interoperable, ensuring the traceability of the entire research process. While there is a range of opportunities of automatizing processes in terms of text labelling by tagging Units of Topography and Actors, there is still a need for trained and experienced historians who decide the level at which data have to be recorded. Anyway, there is a considerable potential in terms of data exploitation and visualization, in which ICT in general should definitely contribute.

ACKNOWLEDGMENTS

This study is part of the current research tasks carried out by the Medieval and Postmedieval Archaeology Research Group GRAMP-UB (2017-SGR-833-GRC) at the University of Barcelona, to which the authors belong, and it is included in our Landscape Archaeology research line. Theoretical development, conceptualization, and applied methods are part of funded research projects developed by the Centre d'Estudis Martorellencs (CLT009/18/00036 – DGPC/exp. 35) and the University of Barcelona (GREYWARE PID2019-103896RJ-I00). The authors wish to thank Ms Noemí Travé for language edition and review.

REFERENCES

- [1] J. Arostegui. *La investigación histórica: teoría y método (Historical research: Theory and method)*, Barcelona, Spain: Crítica, 1995.
- [2] A. Torre. “Un «tournant spatial» en histoire? Paysages, regards, ressources (A spatial turn in history? Landscapes, visions, resources)”, *Annales. Histoire, Sciences Sociales*, vol. 2008, no. 5, pp. 1127-1144, 2018, [On-line] Available: https://www.cairn-int.info/article-E_ANNA_635_1127--a-spatial-turn-in-history.htm
- [3] R. White. *What is Spatial History? Spatial History Lab: Working Paper*, Stanford, USA: Stanford University, 2010, [On-line] Available: <https://web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29>
- [4] D. Cosgrove. “Landscape and Landschaft. Lecture delivered at the «Spatial Turn in History» Symposium”, *German Historical Institute Bulletin*, vol. 35 no. Fall, pp. 57-71, 2004. [On-line] Available: <https://is.cuni.cz/studium/predmety/index.php?do=download&did=130419&kod=JMB067>
- [5] F. Trentmann. “Materiality in the Future of History: Things, Practices and Politics”, *Journal of British Studies*, vol. 48, pp. 283-307, 2009, doi: 10.1086/596123.
- [6] D. Hicks. “The material-cultural turn: event and effect” in *The Oxford Handbook of Material Culture Studies*, Hicks, D.; Beaudry, M. C. Eds. Oxford, UK: Oxford University Press, 2010, pp. 25-98. doi: 10.1093/oxfordhb/9780199218714.013.0002.
- [7] E. Hinrichs, M. Hinrichs, S. Kübler, T. Trippel. “Language technology for digital humanities: introduction to the special issue”, *Language Resources & Evaluation*, vol. 53, pp. 559-563, 2019.
- [8] D.W. Oard. “A Whirlwind Tour of Automated Language Processing for the Humanities and Social Sciences” in *Working Together or Apart: Promoting Digital Scholarship*, Washington DC, USA: Council on Library and Information Resources, 2009.
- [9] J. Kuhn. “Computational text analysis within the Humanities: How to combine working practices from the contributing fields”, *Language Resources & Evaluation*, vol. 53, pp. 565-602, 2019.
- [10] M. Piotrowski. *Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies*, San Rafael, California, USA: Morgan & Claypool Publishers, 2012. [On-line] Available: <https://doi.org/10.2200/S00436ED1V01Y01207HLT017>
- [11] P. Ruiz. *Concept-based and relation-based corpus navigation: applications of natural language processing in digital humanities. Linguistics*. PhD Dissertation. Paris, France: PSL Research University, 2017. [On-line] Available: <https://tel.archives-ouvertes.fr/tel-01575167v2>
- [12] E. Travé, P. Del Fresno, A. Mauri, S. Medina. “The Semantics of historical knowledge. Labelling Strategies for Interdisciplinary and Digital Research in History”, in *HI4NLP 2020 Hybrid Intelligence for Natural Language Processing Tasks 2020. Proceedings of the Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020) Santiago de Compostela, Spain, August 29, 2020*, P. Gamallo, M. García, P. Martín-Rodilla, M. Pereira-Fariña, Eds. Santiago de Compostela, Spain: CEUR, 2020, pp. 17-21, [On-line] Available: <http://ceur-ws.org/Vol-2693/paper2.pdf>
- [13] J. Delve, R. G. Healey. “Is there a role for data warehousing technology in historical research?” in *Humanities, Computers and Cultural Heritage. Proceedings of the XVI International Conference of the Association of Arts and Sciences*, Royal Netherlands Academy of Arts and Sciences Ed. Amsterdam, Netherlands: KNAW, 2005, pp. 106-111. [On-line] Available: <https://dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANShumanitiescomputersandculturalheritageUK.pdf>
- [14] A. Ivanovs, A. Varfolomeyev. “Editing an exploratory analysis of medieval documents by means of XML technologies” in *Humanities, Computers and Cultural Heritage*, Netherlands: KNAW, 2005, pp. 155-160.
- [15] F. de Jong, H. Rode, D. Hiemstra. “Temporal language models for the disclosure of historical text”, in *Humanities, Computers and Cultural Heritage*, Netherlands: KNAW, 2005, pp. 161-168.
- [16] S. Eriksson, H. Skånes. “Addressing semantics and historical data heterogeneities in cross-temporal landscape analysis”, *Agriculture, Ecosystems and Environment*, vol. 139, pp. 516-521, 2010, doi: 10.1016/j.agee.2010.09.011.
- [17] G. Nagypál. “History ontology building: The technical view”, in *Humanities, Computers and Cultural Heritage*, Netherlands: KNAW, 2005, pp. 207-214.
- [18] B. Tomaszewski. “Local Model Semantics, Categories, and External Representation: Towards a Model for Geo-historical Context”, in *Modeling and Using Context. 6th International and Interdisciplinary Conference, CONTEXT 2007, Roskilde, Denmark, August 20-24, 2007, Proceedings. Lecture Notes in Artificial Intelligence*, vol. 4635, B. Kokinov, D. C. Richardson, Th. R. Roth-Berghofer, L. Vieu, Eds. Berlin / Heidelberg, Germany: Springer, 2007, pp. [On-line] Available: <http://context-07.ruc.dk/Context2007DocCons.pdf>
- [19] H. Berger. ‘Microhistory and quantitative data analysis’, In *Humanities, Computers and Cultural Heritage*, Netherlands: KNAW, 2005, pp. 39-42.
- [20] A. Antonacopoulos, D. Karatzas. “Semantics-Based Content Extraction in Typewritten Historical Documents”, in *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR2005)*, IEEE Computer Society Ed. Washington DC, United States: IEEE Digital

- Library, 2005, pp. 48-53, doi: 10.1109/ICDAR.2005.215.
- [21] M. Heller, G. Vogeler. "Modern information retrieval technology for historical documents", in *Humanities, Computers and Cultural Heritage*, Netherlands: KNAW, 2005, pp. 143-148.
- [22] C. Coffin. "The voices of History: Theorising the interpersonal semantics of historical discourses", *Text*, vol. 22, no. 4, pp. 503-508, 2002, doi: 10.1515/text.2002.020.
- [23] G. Nagypál, B. Motik. "A Fuzzy Model for Representing Uncertain, Subjective, and Vague Temporal Knowledge in Ontologies", in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science*, vol. 2888, R. Meersman, Z. Tari, D. C. Schmidt, Eds. Berlin / Heidelberg, Germany: Springer, 2003, pp. 906-923, doi: 10.1007/978-3-540-39964-3_57.
- [24] M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, vol. 3, paper 160018, 2016, doi: 10.1038/sdata.2016.18
- [25] G. Gänser, G. Michetti. *InterPARES Trust Project. Ontology of Functional Activities for Archival Systems*; TR05-Final Report-20180526, 2018. [On-line] Available: <https://interparestrust.org>
- [26] Parthenos. *Guidelines to FAIRify Data Management and Make Data Reusable*. [On-line] Available: <https://zenodo.org/record/2668479>
- [27] A. Mosca. "Ontology-mediated Data Integration and Access in Research and Innovation Policy", in *Proceedings of the Joint Ontology Workshops 2017 (JOWO 2017), Episode 3: The Tyrolean Autumn of Ontology, Bozen-Bolzano, Italy, 21-23 September 2017*, S. Borgo, O. Kutz, F. Loebe, F. Neuhaus, Eds. Bozen-Bolzano, Italy: CEUR, 2017, [On-line] Available: <http://ceur-ws.org/Vol-2050/invited4-Mosca.pdf>
- [28] K. Thibodeau, "The Construction of the Past: Towards a Theory for Knowing the Past", *Information*, vol. 10, paper 332, 2019, doi: 10.3390/info10110332.
- [29] A. Mauri. *La configuració del paisatge medieval: el comtat de Barcelona fins al segle XI (The formation of medieval landscape: The county of Barcelona until the 11th Cent)*. PhD Dissertation. Barcelona, Spain: Universitat de Barcelona, 2006. [On-line] Available: <https://www.tesisenred.net/handle/10803/2065>
- [30] E. C. Harris. *Principles of archaeological stratigraphy (Second edition)*, San Diego, USA: Academic Press, 1989, [On-line] Available: <https://www.sciencedirect.com/book/9780123266514/principles-of-archaeological-stratigraphy>
- [31] A. Carandini. *Historias en la Tierra: Manual de Excavación Arqueológica (Stories from the earth. Excavation Handbook)*, Barcelona, Spain: Crítica, 1997.
- [32] *Foundation Act of the Priory of Sant Genís de Rocafort*. Barcelonense Diocesan Archive, Vol. *Dotialiarum*, no. 5, pp. 31-32.
- [33] E. Mallorquí, *El monestir de Sant Miquel de Cruïlles a l'edat mitjana. Estudi i edició dels documents. 1035-1300 (The monastery of Sant Miquel de Cruïlles in the middle age. Study and edition of documents. 1035-1300)*, Girona, Spain: Ajuntament de la Bisbal d'Empordà, 2000.
- [34] J. Baucells. *El Priorat de Sant Genís de Rocafort (Martorell) (The Priory of Sant Genís de Rocafort)*, Centre d'Estudis Martorellencs. Ed. Martorell, Spain: CEM, 2007.
- [35] J. Salvany. *Sant Genís de Rocafort*. Biblioteca de Catalunya Ed. Barcelona, Spain: BC, 1915.
- [36] J. Salvany. *Sant Genís de Rocafort*. Biblioteca de Catalunya Ed. Barcelona, Spain: BC, 1912.
- [37] I. Hodder, C. Orton. *Spatial Analysis in Archaeology*, Cambridge, UK: Cambridge University Press, 1976.
- [38] *Sant Genís de Rocafort. Ortphoto*. Provided by Institut Cartogràfic i Geològic de Catalunya. Barcelona. [On-line] Available: <http://www.icc.cat/vissir3/>
- [39] C. González-Pérez. "A Conceptual Modelling Language for the Humanities and Social Sciences", in *Sixth International Conference on Research Challenges in Information Science (RCIS)*, IEEE Computer Society Ed. Washington DC, United States: IEEE Digital Library, 2012, doi: 10.1109/RCIS.2012.6240430.
- [40] E. Travé, P. Del Fresno, A. Mauri. "Ontology-mediated Historical Data Modelling: Theoretical and Practical Tools for an Integrated Construction of the Past", in *Digital Humanities Special Issue*. C. González-Pérez Ed. *Information*, vol. 11, paper 182, doi:10.3390/info11040182.
- [41] P. Del Fresno. *Sistema de Información Arqueológica: propuesta de normalización, desarrollo conceptual e informático. (Archaeological Information System: Normalization proposal, conceptual and computing development) PhD Dissertation*. Vitoria, Spain: University of the Basque Country, 2016. [On-line] Available: https://www.academia.edu/23836202/SigArq_Sistema_de_Información_Geospacial_Arqueológica. [On-line] Available: <http://www.sigarq.es/home/>
- [42] H. Wickham, 'Tidy Data', *Journal of Statistical Software*, vol. 59, no.10, pp. 1-10, 2004, doi: 10.18637/jss.v059.i10.
- [43] A. Mauri, E. Travé, P. Del Fresno. "An integrated implementation of written and material sources – conceptual challenge and technological resources", in *Archaeology. New Approaches in Theory and Techniques*, I. Ollich Ed. Rijeka, Croatia: InTech, 2012, pp. 41-64, doi:10.5772/38934.
- [44] P. Del Fresno, A. Mauri, E. Travé. "Documentació d'intervencions de restauració en jaciments arqueològics i patrimoni edificat: cap a la construcció d'un Sistema d'informació (Recording restoration tasks in archaeological sites and building heritage: towards the construction of an information system)". *Documentar en la conservació-restauració. Mètodes i noves technologies*, Conservadors Restauradors Associats de Catalunya. Barcelona: Spain, CRAC, 2020, in press.
- [45] C. M. Sperberg-McQueen, L. Burnard. *TEI P5: Guidelines for Electronic Text Encoding and Interchange TEI*. Text Encoding Initiative Consortium, 2019. [On-line] Available: <https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>
- [46] E. Travé. "Los hornos medievales y la cerámica negra" in *Actas del IX Congreso de la Asociación de Ceramología. Obra negra y alfarería de cocina*, Asociación de Ceramología Ed., Quart, Spain: Ajuntament de Quart, 2017, pp. 123-147. [On-line] Available: <http://www.ceramologia.org/gestion/archivos/1084.Est.pdf>
- [47] R. Sprugnoli, S. Tonelli. "One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective", *Natural language engineering*, vol. 23, no. 4, pp. 485-503, 2016, doi: 10.1017/S1351324916000292.



Esther Travé Allepuz

Esther Travé holds a Degree in History (Universitat de Barcelona, Spain, 2005) and a PhD in Medieval History at the same university (2009). She is a medieval archaeologist and her research field is the production and distribution of greyware pottery in medieval and postmedieval ages. She was trained in Ceramic Petrography at the University of Sheffield (2008) and in XRF and XRD techniques at the Institute of Archaeology, University College London (UCL) (2012-2014). She lectures at the University of Barcelona since 2015, and she currently is leading researcher of the archaeological project developed at the Sites of Santa Margarida and Sant Genís de Rocafort (Martorell). Her most recent research has been devoted to data analysis and the development of interdisciplinary approaches applied, in particular, to the historical and archaeological study of pottery.



Pablo del Fresno Bernal

Pablo Del Fresno obtained his Degree in History at the University of the Basque Country (Spain, 2000). He holds an MSc in Geographic Information Technologies (Universitat Autònoma de Barcelona, Spain, 2008) and a PhD in History (University of the Basque Country, 2016). He works as a professional archaeologist at Sistemes de Gestió de Patrimoni SCCL. Based upon the theoretical reflection of his PhD dissertation, he is the development coordinator of the Archaeological Information System SigArq, within a long collaborative framework in which the Geographic Information and Remote Sensing Lab (LIGIT) at the UAB takes part. He participates in several archaeological projects in which the SigArq system is involved. Amongst these, Santa Catalina (Mansilla de la Sierra, La Rioja), the Convent de Penyafort (Santa Margarida i els Monjos, Barcelona), and the Archaeological Research Project of Santa Margarida and Sant Genís de Rocafort (Martorell, Barcelona), are the more relevant.



Sonia Medina Gordo

Sonia Medina is a PhD candidate working under Esther Travé's supervision at the University of Barcelona. Her PhD project aims at developing digital strategies for data management and exploitation and applying them to the historical analysis of the formation of the Castilian Feudal System at the Upper Arlanza Basin (Burgos, Spain). She holds a Degree in History (Universitat de Barcelona, Spain, 2018) and a MA in Advanced History Studies (Universidad de Salamanca, Spain, 2019). Approaching the medieval period according to an interdisciplinary method integrating results from historical science and archaeology led her to join this research team. She contributes to the data modelling and database development, the way of dealing with uncertainty and bias in historical discourses being one of her main interests.



Alfred Mauri Martí

Alfred Mauri has a long career in Archival Science and Archaeology since he obtained his Degree in Prehistory, Ancient History and Archaeology (Universitat de Barcelona, Spain, 1983). From 1981 to 2002 he worked in Martorell (Spain), as an archivist in charge of the Town Council Archive, while he was the leading researcher at the Archaeological Site of Santa Margarida, located in the same town. He performed his PhD Dissertation in the field of Landscape Archaeology and developed the concept of Unit of Topography. After obtaining his PhD (Universitat de Barcelona, Spain, 2006) he has been Lecturer and Degree coordinator at the High School of Archival Science and Records Management of the Universitat Autònoma de Barcelona. He has been director of this same centre for three years (2016 – 2018). His main research lines are landscape archaeology and archaeology of architecture, the integrated and georeferenced historical data management, and Big Data analysis and exploitation.