

This is a postprint version of the following published document:

GARCÍA ZORITA, C. & PACIOS, A. R. (2018). Topic modelling characterisation of Mudejar art based on document titles. *Digital Scholarship in the Humanities*, Volume 33, Issue 3, pp. 529–539.
<https://doi.org/10.1093/lc/fqx055>

Topic modelling characterisation of Mudéjar art based on document titles

Carlos Garcia-Zorita and Ana R. Pacios

Departamento de Biblioteconomía y Documentación, Universidad Carlos III de Madrid, Spain

Abstract

Text mining techniques were applied to a corpus consisting in the titles of 2454 documents on Mudejar art, a style unique to Spanish art history. Probabilistic topic modelling was used to analyse the semantic structure underlying the suite of documents studied. Two classifications were obtained, an initial, generic division into five topics followed by a second more refined division into ten. These were compared to the preliminary subject categories found for the corpus with the guidance of an area specialist. The classifications delivered by the automatic and manual procedures were observed to be compatible. The conclusion drawn was that the deployment of digitised data affords the opportunity to conduct humanities studies from new perspectives.

Keywords: Probabilistic topic modelling, Mudejar art, bibliographic repertories, publication titles.

Introduction

The onset of digital technology applied to the humanities is bringing about a paradigm change in knowledge generation and problem solving in these disciplines. The 2004 publication of *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens and John Unsworth (27) prompted reflection on the nature and characteristics of the digital humanities and encouragement of their use by humanities scholars (Weingart). With the delivery of more dynamic and visual research, they have made their impact felt for over a decade.

Digital humanities studies are associated, among others, with the creation of data repositories, the application of processing methodologies and the generation of new knowledge that favours the advance of humanistic research. The discipline is, in short, the result of dynamic dialogue between technology and humanistic research. Data processing of this type is not feasible without the interdisciplinary collaboration of IT engineers, data analysts, bibliometricians (Pacios-Lozano and García-Zorita, 2016a, 2016b) and, of course, humanists. The benefits for disciplines such as library science and digital humanities, as well as for the scientists involved in such research, have been highlighted by a number of authors (Robinson, Priego and Bawden). The necessary collaboration and teamwork entailed have introduced a significant and steady change in humanists' practice, as paper authorship shows (Wuchty et al.). Some authors, such as Bosch, see the digital humanities as a space for practice and experience that favour the convergence of document processing, the theoretical rapprochement characteristic of the objects of humanistic study and digital methodology and tools. Dobrecky contends that their consolidation in scientific and academic domains has an international impact.

Cita:

GARCÍA ZORITA, C. & PACIOS, A. R. (2018). Topic modelling characterisation of Mudejar art based on document titles. *Digital Scholarship in the Humanities*, Volume 33, Issue 3, pp. 529–539. <https://doi.org/10.1093/lhc/fqx055>

The digital humanities may be said to have earned social acknowledgement on the grounds of the factors that, according to Liu, are pooled under the heading: a critical mass of researchers, papers, scientific meetings, institutionalisation (institutes, academic curricula, associations, job offerings) and notorious visibility. That, in conjunction with the epistemological endorsement of research conducted with scientific methods that has yielded applicable results, confirms the scientific nature of the digital humanities and their identification as an emerging discipline.

The digital humanities are vested with scientific discipline status, among others, by the use of methodological tools related to ‘text mining’, in particular the technique known as ‘probabilistic topic modelling’ (PTM) deployed in this study. The collaboration between humanists and IT engineers and statisticians that has spawned these techniques entailed the development of model components and algorithms adapted to humanistic topics and texts (Blei). PTM is based on a suite of algorithms able to detect and extract latent semantic relationships from large text-based corpora. The earliest applications in the field date from 2002, with interest intensifying steeply after 2010 (Meeks and Scott). Knowledge can be extracted with the technique through the analysis of text, data and any underlying relationships among them (Arcila, Barbosa and Cabezuelo-Lorenzo).

The chief output of topic modelling, in particular of the techniques using the latent Dirichlet allocation (LDA) algorithm introduced by Blei et al., is the generation of series of words that tend to appear together with other words in the same contexts (individual documents) throughout the entire corpus. These words are assigned to categories that represent ‘latent topics’ (Carron-Arthur, Reynolds, Bennett, Bennett and Griffiths). Blei defined a topic to be ‘a distribution over a fixed vocabulary’. Each category or topic, sometimes referred to as ‘bags of words’ must be interpreted by the researcher to vest them with meaning in the context of the corpus. Each document is also assigned to every topic extracted by the algorithm with a variable probability. If a probability threshold is established, each document can be assigned to the topic or topics with which it exhibits the greatest semantic affinity. Other suitable methods can be consulted in Newman and Block.

This study used PTM to attempt to verify the utility and validity of this methodology for the characterisation of an artistic style unique in Spanish art history: Mudejar art.

It has taken a century and a half of scholarship to reach today’s accepted characterisation and definition of Mudejar art, one of the most controversial styles in the history of Spanish art. That controversy has often stemmed from an interpretation prone to reduce the weight of Islam and Muslim tradition, fruit in turn of a mis-read of the legacy of Al-Andalus in Spanish culture and its attribution of a merely ornamental role. From its introduction as an artistic style in José Amador de los Ríos’s famous (and famously challenged) speech in 1859 through the present, there has been no want of interpretations that denied its artistic reality. Its singularity, the outcome of a multiple and diverse, dynamic and rich historic-artistic era, the study of which is imperative to understanding the factors that made it possible, favoured the application of qualifiers such as a ‘typically Hispanic cultural epiphenomenon’ (Valdés 84).

Although Mudejar art has been analysed primarily by Spanish researchers, interest in the style has crossed the bounds of the Iberian Peninsula to become an object of university and institutional study in other European countries as well as in America. The resulting vast historiography was explored from a historic-bibliometric perspective in two consecutive studies. The first addressed the factors relating to the distribution over time, document typologies and journals (Pacios Lozano and García Zorita, 2016a) and the second the languages in which the studies were published, subject matters, authorship and distribution of articles by author (Pacios-Lozano and García-Zorita, 2016b). The raw material for these studies was a suite of 2454 bibliographic records on papers published between 1857 and 2013 obtained from three bibliographic repertories on Mudejar art compiled at different times (Pacios, 1993; Pacios, 2002 and Pacios, unpublished).

The originality of this article lies in the application of topic modelling to art history and the use as a corpus of the titles of the documents referenced, rather than the full documents or abstracts, further to standard practice.

Purpose and method

The study used probabilistic topic modelling, a text mining technique, to identify the key features (characteristics, attributes) of Mudejar art studied by art historians. As Gretarsson et al. (5) noted, PTM is ‘a widely-used unsupervised machine learning technique for automatically extracting semantic or thematic topics from a collection of text. The topics provide a high-level abstract representation of documents in a corpus, and can be used for searching, categorizing, and navigating through collections of documents’.

All the elements recommended by Brett as requirements to work with topic modelling were available in this endeavour.

- *A corpus*, preferably with over 1000 items. In this case the corpus consisted in the titles of the 2454 bibliographic records alluded to earlier. Since the bibliographies contained an analytical abstract of the document (journal articles, monographs, congress papers) associated with each record, the possibility of processing those texts was initially contemplated. Inasmuch as the abstracts were drafted by the author of the repertories, however, they were deemed to be less objective than the document titles as defined by their authors. An earlier ad hoc subject classification of Mudejar art conducted by the authors (Pacios-Lozano and García-Zorita, 2016a) based on the content of each document and their knowledge of the artistic style was the material this exercise aimed to verify via text mining and unsupervised, automatic classification.
- *Familiarity with the corpus*. Brett contended that the only way to know whether the results of topic modelling make sense or at least are not ‘wildly off’ is to have some knowledge of what is being sought. In the present case, that knowledge was ensured by previous studies and one of the authors’ PhD.-substantiated expertise in the area.
- *A tool to do the topic modeling*. This study applied MALLET (McCallum), one of the most popular and widely used digital humanities tools, which runs on the aforementioned LDA algorithm.
- *A way to understand results*. A two-phase analysis was deployed to obtain the results. An initial attempt was made to obtain a general classification consistent with the semantic structure of the Mudejar art corpus by selecting five topics,

whilst in a second phase, in which a more specific contextual analysis was pursued, ten topics were selected. Inasmuch as the documentary corpus was in Spanish, that language's stop-words (list of words with no semantic meaning) were used in the first phase. The corpus was pre-processed to delete accents and change upper to lower case letters. Neutral words that added no semantic value to the corpus were eliminated in the second phase; that entailed deleting elements such as century/ies and the respective Roman numerals used in Spanish to designate them (XI, XII); art, artistic, historic, national, Hispanic; and Saint, St and saint's names: Mary, Michael, Peter and so on.

Interpreting the results obtained with MALLET was neither simple nor straightforward, an experience reported by other authors (Navarro Colorado), particularly where researchers were unfamiliar with the subject at issue. In this case, the interpretation of the results was 'unsupervised', i.e., the algorithm was not subject to decisions mandated by researcher-defined reference features (Dobson 548). Therefore, no prior condition was established that would enable the algorithm to unambiguously assign each text or group of words considered to a given category.

Results

The assignment of the words in the titles of the documents on Mudejar art to a small number of topics that revealed the latent general structure of the corpus is discussed first. The five topics obtained, together with the ten most probable words and phrases in each, are given in Table 1. Each topic was labelled with a short descriptor. The table lists the terms delivered by the algorithm in the original Spanish followed by the English translation in round brackets, although the proper nouns in monument names such as 'Santa María' or 'Santa Clara', the names of monumental buildings such as 'El Alcázar' or 'La Alhambra', place names such as Sevilla or Toledo, regions or Mudejar art 'hubs' such as Aragón or Andalucía and geographic adjectives such as 'aragonés' or valenciano have been left untranslated. Some of the remarks contain references to terms [shown in square brackets] that are not among the most likely but are needed to support the observations.

Table 1: List of five topics extracted after analysis

Topic #	Identifier	10 top-ranking words & phrases	Remarks
Topic 0	Religious architecture	iglesia parroquial (parish church), SANTA MARÍA, SANTA CLARA, NUESTRA SEÑORA, mudéjar (Mudejar), monasterio (monastery), torre (tower), ZARAGOZA, catedral (cathedral), TERUEL	This topic was identified on the grounds of terms related to architectural types (parish church, tower, monastery, cathedral), supplemented by saints' names associated with monuments (Holy Mary, Saint Clair, Our Lady, [Saint Dominic], [Peter the Martyr]).
Topic 1	Mudejar art or Mudejar style	arte mudéjar (Mudejar art), arquitectura (architecture), mudéjar, cerámica (ceramics), mudéjares (Mudejar), ARAGÓN, medieval (Medieval),	Here the terms were related to characteristics of [Mudejar style], the types of artistic expression in which it is most visible (architecture, ceramics, [joinery], even [town planning]), the (Mudejar) craftsmen, the geographic

Cita:

GARCÍA ZORITA, C. & PACIOS, A. R. (2018). Topic modelling characterisation of Mudejar art based on document titles. *Digital Scholarship in the Humanities*, Volume 33, Issue 3, pp. 529–539. <https://doi.org/10.1093/lhc/fqx055>

Topic #	Identifier	10 top-ranking words & phrases	Remarks
		provincia (province), ciudad (city), ARAGONÉS, TOLEDO	areas where the style is most densely represented (Aragon, Toledo), and the attribution of elements primarily to the Middle Ages.
Topic 2	Chronology	siglo (century), <i>XVI</i> , siglos (centuries), arquitectura (architecture), madera (wood), mudéjares (Mudejar), iglesias (churches), XIII, XVII, ZARAGOZA	This topic included words associated with the (thirteenth, [fourteenth], [fifteenth], sixteenth, seventeenth and [eighteenth]) centuries when the style was applied, and in particular with the key manifestations: architecture and roof frames/ceilings.
Topic 3	Joinery	Carpintería (joinery), CASTILLA, media (middle), LEÓN, edad (age/s), blanco (pinewood), mudéjar (Mudejar), alarifes (joiners), tratado (treatise), armar (to join)	This topic is identified with terms related to (pinewood) joinery, its craftsmen, the joiners, and the treatise authored by [Diego López] that served as a basis for its evolution.
Topic 4	Civil architecture	palacio (palace), mudéjares (Mudejar), SEVILLA, casa (house), alcázar (royal palace-fortress), real (royal), TOLEDO, Pedro, restauración (restoration), museo (museum)	Here the terminology alluded reiteratively to the various types of civil architecture (royal palace-fortress, palace, house, [castle]) and the royal nature of some works built under the initiative of the Catholic Monarchs or [Pedro I]. This architecture is characterised by the rich indoor [plasterwork] ornamentation with [inscriptions] and [mural paintings]. Some monuments, as in Toledo and Seville, have been turned into museums.

This initial classification of the documents in the corpus referred to more general matters including, obviously, Mudejar style itself (Topic 1), under which a substantial portion of the documents could be classified; its chronology (Topic 2); the realms in which it is primarily found: civil (Topic 4) and religious (Topic 0) architecture; and one of its most prominent and widely studied artistic manifestations: joinery (Topic 3). In addition, each and every one of the five topics contained scattered references to terms related to the highly territorial nature of the style and in particular to the regions of Spain with the densest presence (known in the specialised literature as ‘Mudejar hubs’) of artistic manifestations, such as Aragon, Toledo and Castile and specific provinces in those regions, including Zaragoza, Teruel, Toledo, Leon and Seville.

In a second phase, a more specific and refined classification established to better achieve the objective identified 10 topics under which to classify the corpus documents. The results are given in Table 2.

Table 2: List of 10 topics extracted after analysis

Topic #	Identifier	10 top-ranking words & phrases	Remarks
T0	Seville's Alcazar	palacio (palace), SEVILLA, ALCÁZAR, real (royal), Pedro, reyes (kings), restauración (restoration), armadura (roof frame/ ceiling), ALHAMBRA, reales (royal)	This topic revealed the weight of the studies on Seville's Royal Alcazar, one of the compounds that best represents the style, of which Pedro I's palace, a masterpiece of Andalusian Mudejar construction, forms part. Its ornamental [ceilings] are the identifying element. Royal patronage, in particular by the [Catholic Monarchs] and Pedro I, favoured the construction of works of enormous interest.
T1	Aragonese religious architecture	iglesia parroquial (parish church), ZARAGOZA, mudéjar (Mudejar), torre (tower), TERUEL, catedral (cathedral), CALATAYUD, techumbre (roof), torres (towers), Pedro	The topic identified religious Mudejar architecture in Aragón. Free-standing or church-attached towers are among the most characteristic elements of Aragonese Mudejar architecture. Examples may be found at Calatayud, Zaragoza and Teruel. Many parish churches in Aragon have a singular structure and floor plan, as well as towers that on occasion drew inspiration from [minarets]. Some have wooden roofs.
T2	Andalusian religious architecture	Arquitectura mudéjar (Mudejar architecture), provincial (provincial), GRANADA, iglesias (churches), religiosa (religious), SEVILLA, civil (civil), ladrillo (brick), ciudad (city)	The terms identified alluded to religious architecture in Andalusia, with representative examples in Granada and Seville. Reference was also made to the Mudejar building material par excellence, brick.
T3	Territorial (local/ regional) nature	Mudéjar (Mudejar), ARAGONÉS, islámico (Muslim), TOLEDANO, mudéjarismo (Mudejarism), castillo (castle), ANDALUCÍA, técnicas (techniques), ISLAS CANARIAS, morisco (Moorish)	The terms of this topic revealed the highly local nature of the style, its singularity and variety from region to region and school to school (Aragon, Toledo, Andalusia, Canaries, Castile). Its persistence beginning in the sixteenth century in what is known as 'Moorish' [early Renaissance] style was also addressed. References to [materials], [techniques] and [ornamentation] attested to studies on the ongoing presence of the [Muslim world].
T4	Muslim tradition	CASTILLA, media (middle), LEÓN, arquitectura (architecture), TOLEDO, baja (early), ANDALUS, BURGOS, tradición (tradition), CÓRDOBA, sinagoga (synagogue), ermita (shrine)	The legacy of Al-Andalus is present in other Mudejar art hubs. In particular, the terms in this topic revealed its presence in the architecture of the [former reigns] of Castile and Leon and Toledo. Synagogues and [mosques] converted to [shrines] contributed to the persistence of [Muslim tradition].
T5	Wood	madera (wood), mudéjares (Mudejar),	Wood, the material used in the structures that roof many Mudejar buildings such as churches,

Topic #	Identifier	10 top-ranking words & phrases	Remarks
		iglesias (churches), armaduras (roof frames/ceilings), cubiertas (roofs), arcos (arches), provincial (provincial), PALENCIA, VALENCIANA, sur (south)	was one of the building materials most representative of the style.
T6	Aragonese civil architecture	ARAGÓN, decoración (decoration), casas (houses), luna ('luna'), moros (Moors), ARAGONESES, azulejos (tiles), palacios (palaces), moriscos (Moorish), torres (towers), TOLEDO, ARAGÓN, artes (arts), arqueológico (archaeological), techumbres (roofs), artesonados (coffered ceilings)	This topic included terms referring to Aragonese civil architecture. Mudejar [Moor] construction of some towers and palaces, such as the one belonging to 'Papa Luna' (a fourteenth century 'dissident' pope), is documented. Some of these structures have richly decorated coffered ceilings.
T7	Ceramic decoration	cerámica (ceramics), TOLEDO, ermita (shrine), castillo (castle), decoración (decoration), cristo (Christ), influencia (influence), restauración (restoration), arquitectónica (architectural), islam (Islam), MURCIA, hispanomusulmana (Hispano-Muslim), bajomedieval (early Middle Ages), ARAGONESA, PATERNA, VALENCIANA	Here the terms referred to Mudejar decoration, in particular tile. Many excellent examples of Muslim decorative elements can be found. One is pottery bearing [lustre] or [metallic glints], the most prominent production sites of which were Valencia, [Murcia], Paterna and [Aragon].
T8	Monastic and convent architecture	mudéjares (Mudejar), TOLEDO, monasterio (monastery), iglesia (church), casa (house), real (royal), CLARA, convento (convent), sinagoga (synagogue), capilla (chapel), CÓRDOBA, SANTIAGO,	Here the top-ranking words and phrases referred to the architecture of religious orders, monasteries and convents, as well as the royal chapels they housed, with a sizeable decorative repertory of interest and representative examples at Toledo, Cordoba and Guadalajara The synagogue, another architectural structure of great ornamental beauty reused in the Christian world, such as the one at Segovia, was also present.

Cita:

GARCÍA ZORITA, C. & PACIOS, A. R. (2018). Topic modelling characterisation of Mudejar art based on document titles. *Digital Scholarship in the Humanities*, Volume 33, Issue 3, pp. 529–539. <https://doi.org/10.1093/llc/fqx055>

Topic #	Identifier	10 top-ranking words & phrases	Remarks
		GUADALAJARA, SEGOVIA	
	Treaty on joinery	carpintería (joinery), blanco (pinewood), alarife (joiner), tratado (treatise), geometría (geometry), compendio (compendium), Arenas, armar (to join), Lopez, Diego	All the words in this topic were associated with Diego López de Arenas's treatise on pinewood joinery, applied by joiners. The [geometric lazo] was a constant presence in the context of roof frame decoration.

The characterisation of Mudejar art afforded by the aforementioned 10 topics distinguished among elements, firstly, on the grounds of their geographic location. That was consistent with the geographic or territorial approach to research on this artistic style and the bibliographies used as the point of departure for this study. In particular, the Mudejar hubs with the densest presence of artistic elements were widely represented: Aragon, Castile, Toledo and Andalusia. That was particularly visible in T3, labelled Territorial (local/regional) nature, although provincial and city place names particularly significant for a specific manifestation or their wealth of Mudejar art appeared in all the topics: Leon, Burgos, Zaragoza, Calatayud, Canarias, Teruel, Cordoba, Seville, Granada, Palencia, Murcia, Valencia, Segovia and Guadalajara.

The various architectural typologies (T0 and T4) were also prominent, along with joinery, one of the chief artistic expressions of the style. Some topics focused on highly significant manifestations, such as Seville's Alcazar (T0). Another topic (T9) revolved around Diego López de Arenas's sixteenth century treatise on joinery that served as a basis for the construction of many of the wooden roofs so characteristic of the style. T7, in turn, was associated with the presence of Mudejar decoration, another logical result, given the eminently decorative nature of the style.

In light of their specificity, these ten topics could be clearly linked to the initial five generic categories. The resulting relationships and hierarchies are listed in Table 3.

Table 3: Relationships between the two unsupervised classifications

Topic 0	Religious architecture	
	T8	Monastic and convent architecture
	T1	Aragonese religious architecture
	T2	Andalusian religious architecture
Topic 1	Mudejar art or style	
	T3	Territorial (local/regional) nature
	T4	Muslim tradition
	T5	Wood
	T7	Ceramic decoration
Topic 2	Chronology	
Topic 3	Joinery	
	T9	Treatise on joinery
Topic 4	Civil architecture	
	T6	Aragonese civil architecture
	T0	Seville's Alcazar

The list in Table 3 shows that the objects of study and priority attention in Mudejar art identified in the five initial topics were refined and defined more clearly by the ten topics comprising the second classification. The characteristics of *Mudejar art or style* (Topic 1) included its territorial (local/regional) nature (T3), mirroring its diversity; the power of Muslim tradition (T4), fruit of the persistence of Hispano-Muslim art after the Christian reconquest of Al-Andalus; wood (T5) as one of the materials most widely used in this type of art, also associated with Hispano-Muslim art; and ceramic decoration (T7), attesting to the continuity of Muslim decoration based on *horror vacui*, with the repetition and reiteration on edging of all manner of themes (plant, epigraphic, geometric). The topic *religious architecture* (Topic 0) was broken down into monastic and convent architecture (T8) and its significance in Aragon (T1) and Andalusia (T2). Similarly, under *civil architecture* (Topic 4) this second process revealed the predominance of Aragonese manifestations (T6) and identified Seville's Alcazar (T0) as the preferred object of monographic studies. The findings for *Topic 3, joinery*, in turn, showed that Diego López de Arenas's treatise (T9) was another of the key subjects explored by Mudejar art researchers.

The relevance of the subjects receiving priority attention from such researchers according to the unsupervised classifications into five and ten topics was checked against an overview of Mudejar historiography (in Spanish) authored by Professor Gonzalo Borrás, a specialist of high repute in Mudejar art.

The analysis was validated by comparing the automatic classifications obtained with unsupervised machine learning to a manual classification of the same corpus of documents under expert guidance. An earlier study (Pacios-Lozano and García-Zorita, 2016a, 83) established 13 categories in Mudejar art studies on the grounds of subject area or type of artistic manifestation. Table 4 lists the relationships among the three classifications, reproducing in the first column (headed 'Expert'), the order determined for the aforementioned subject categories.

Table 4: Expert manual vs unsupervised TPM classification

Expert	Unsupervised (five topics)	Unsupervised (ten topics)	Terms related to the three classifications
1. GENERAL	Topic 1 - Mudejar art or style	T3- Territorial (local/regional) nature	Aragonese Toledan Andalusia Canary Islands Castilian province Castile Aragon Palencia Valencian Burgos Leon Cordoba Toledo
		T4- Muslim tradition	synagogue shrine Muslim tradition Al-Andalus
	Topic 2 - Chronology	---	early Medieval Gothic-Mudejar Moorish century numerals (deleted as neutral terms)
2. THEORY ON STYLE, INTERPRETATION, TERMINOLOGY	Topic 1 - Mudejar art or style	T3- Territorial (local/regional) nature	Mudejarism
3. CRITIQUE OF TREATISES	Topic 3 - Joinery	T9- Treatise on joinery	joinery pinewood treatise to join López geometry Arenas compendium Diego geometric lazo points
4. MATERIALS AND TECHNIQUES	Topic 1 - Mudejar art or style	T5- Wood	wood roof frames/ceilings roofs arches structure joinery brick techniques
5. ARCHITECTURE	Topic 0- Religious architecture	T8- Monastic and convent architecture	monastery convent church chapel Tordesillas

Cita:

GARCÍA ZORITA, C. & PACIOS, A. R. (2018). Topic modelling characterisation of Mudejar art based on document titles. *Digital Scholarship in the Humanities*, Volume 33, Issue 3, pp. 529–539. <https://doi.org/10.1093/llc/fqx055>

		T1- Aragonese religious architecture	parish church Zaragoza tower Teruel Cathedral Calatayud cathedral tomb
		T2- Andalusian religious architecture	architecture Granada churches joinery brick religious Seville shrine synagogue
	Topic 4- Civil architecture	T6- Aragonese civil architecture	house houses Aragonese Aragon typology civil Castile
		T0- Seville's Alcázar	palace Seville alcazar courtyard king
6. JOINERY	Topic 3 - Joinery	T9- Treatise on joinery	roof frames/ceilings roofs ceilings coffered ceilings
7. DECORATION	Topic 1 - Mudéjar art or style	T7-- Ceramic decoration	ceramics decoration architectural glint pottery Manises
8. TOWN PLANNING	Topic 4- Civil architecture	---	city houses
9. SUMPTUARY AND DECORATIVE ART	Topic 1 - Mudéjar art or style	T7-- Ceramic decoration	Ceramics book binding
10. RESTORATION AND CONSERVATION	Topic 4- Civil architecture	T0- Seville's Alcazar	palace Seville alcazar (royal palace-fortress) restoration conservation
11. CRAFTSMEN	Topic 3 - Joinery	T9- Treatise on joinery	joiners
12. PHILANTHROPY	Topic 4- Civil architecture	T0- Alcázar at Sevilla	royal Pedro kings
13. OTHER			archaeological

The second (Unsupervised (five topics)) and third (Unsupervised (ten topics)) columns list the classifications defined with topic modelling methodology. The existence of a relationship among the three classifications can be deduced from the use of the same terms (last column in Table 4) in the unsupervised and the manual classification (column headed Expert).

The data in Table 4 attest to non-univocal correspondence between the 13 manual categories and the five topics in the initial automatic classification. The expert associated each manual category with one of the first five topics, with two exceptions: the documents classified manually under the heading GENERAL, which were distributed under Topics 1 (Mudejar art or style) and 2 (Chronology), and the ones under the category ARCHITECTURE, which the automatic classification divided into Civil (Topic 0) and Religious (Topic 4) architecture. The topics identified in the automatic classification embraced several of the manually classified categories: aforementioned Topic 1 also included documents manually classified as THEORY ON STYLE, INTERPRETATION, TERMINOLOGY; MATERIALS AND TECHNIQUES; DECORATION; AND SUMPTUARY AND DECORATIVE ART. Documents on CRITIQUE OF TREATISES; JOINERY; and CRAFTSMEN were classified under Topic 3. Documents on TOWN PLANNING; RESTORATION AND CONSERVATION; and PHILANTHROPY and the aforementioned general category ARCHITECTURE were listed under Topic 4.

A comparison of the automatic classification into 10 topics and the manual division yielded results similar to the findings for the five topic classification, with the exception of Topic 2 (Chronology) which, for reasons of methodological construction, had no equivalent. The very general nature of the category TOWN PLANNING led to the same result.

Of particular interest was the breakdown of the category ARCHITECTURE in the second automatic classification into Monastic and convent architecture (T8), Aragonese religious architecture (T1), Andalusian religious architecture (T2), Aragonese civil architecture (T6) and Seville's Alcázar (T0).

Conclusions

Text mining, a 'topic modelling' technique, has proven to be a valid tool for discovering the invisible subject structure in a corpus of documents. This study showed that it is also useful for analysing an unstructured collection of titles, with each title understood to constitute the minimum information representative of the respective document. Whilst scientific paper titles do not necessarily reflect their content, when their number is large enough, they may be useful to classify a group of documents automatically with unsupervised machine learning. Expert intervention is nonetheless necessary to extract the relationships among these words based on their distribution and to know what to do with the terms so identified.

Comparing the subject classification of the documents delivered by 'topic modelling' to the classification of the same corpus of documents based on expert knowledge (Pacios and García-Zorita, 2016a, 79) provided insight into the validity of these techniques, proven by their capacity to pool and break down categories. The conclusion drawn is that the manual and automatic methods concurred in detecting the following as the most significant features of the Mudejar art studies analysed.

- The two expressions of Mudejar art that attracted most research interest were architecture and joinery.
- The Mudejar hubs on the Iberian Peninsula studied most intensely were Andalusia and Aragon.
- Seville's Alcazar was the object of more monographic studies (conducted from a number of scholarly perspectives) than any other monument.

The present application of 'topic modelling' to a series of titles revealed the potential of the technique, for instance, for library or bookstore catalogue search indexing or filtering. The digital humanities can be used in many lines of research, deploying statistical methods such as in this study or other supplementary techniques to represent or visualise data. All, however, call for the availability of machine-legible and hence processable data and corpora of documents. For certain subject areas such as art history, particularly in territory-specific realms such as Spanish culture, history and art, however, digitisation is unfortunately a distant goal.

References

Amador de los Ríos, José (1859). *Discursos leídos ante la Real Academia de Nobles Artes de San Fernando, en la recepción pública de Don José Amador de los Ríos*. Madrid: Imprenta de José Rodríguez.

Arcila-Calderón, Carlos, Eduar Barbosa-Caro y Francisco Cabezuelo-Lorenzo (2016). "Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística". *El profesional de la información* 25:4, 623-631. <http://dx.doi.org/10.3145/epi.2016.jul.12>

Blei, David M. (2012a). "Topic Modeling and Digital Humanities". *Journal of Digital Humanities* 2.1 <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/>

Blei, David M. (2012b) "Probabilistic topic models". *Communications of the ACM* 55.4 (2012): 77-84. <http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf>

Blei, David M., Andrew Y. Ng. and Michael I. Jordan. (2003). "Latent dirichlet allocation". *Journal of machine Learning research* 3: 993–1022.

Borrás Gualis, Gonzalo M. (1990). *El arte mudéjar*. Teruel: Instituto de Estudios Turolenses.

Bosch, Mela. (2015). "Cóncavo y convexo: Documentación y Humanidades Digitales, punto de inflexión". *Actas de las I Jornadas de Humanidades Digitales*. Lucía Cantamutto, Gimena del Rio Riande, Gabriela Striker, eds. Ciudad Autónoma de Buenos Aires: Editorial de la Facultad de Filosofía y Letras Universidad de Buenos Aires. https://www.researchgate.net/publication/303042134_Concavo_y_convexo_Documentacion_y_Humanidades_Digitales_punto_de_inflexion

Brett, Megan R. (2012). "Topic Modeling: A Basic Introduction". *Journal of Digital Humanities* 2.1 <http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/>

Dobrecky, Leticia P. (2015). "Humanidades Digitales y Bibliotecas: un nuevo desafío". *47a Reunión Nacional de Bibliotecarios (ABGRA)*. Ciudad Autónoma de Buenos Aires (Argentina), 21-23 Abril. (Unpublished) [Conference poster] <http://eprints.relis.org/25549/>

Dobson, James E. (2015). "Can an algorithm be disturbed? Machine learning, intrinsic criticism, and the Digital Humanities". *College Literature: a journal of critical literary studies* 42.4, 543-564.

Gretarsson, Brynjar, John O'Donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David and Smyth Newman, Padhraic (2012). "TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling". *ACM Transactions on Intelligent Systems and Technology* 3.2, doi>[10.1145/2089094.2089099](https://doi.org/10.1145/2089094.2089099)

Liu, A. (2013). "The meaning of the Digital Humanities". *PMLA (Publications of the Modern Language Association of America)* 128.2: 409-23.

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. <http://mallet.cs.umass.edu>

Meeks, Elijah and Scott B. Weingart (2012). "The Digital Humanities Contribution to Topic Modeling" *Journal of Digital Humanities* 2.1 <http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/>

Newman, David J., and Sharon Block (2006). "Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology*, 57.6: 753-767. <https://doi.org/10.1002/asi.20342>

Pacios Lozano, Ana R. (1993). *Bibliografía de arquitectura y techumbres mudéjares (1857-1991)*. Teruel: Instituto de Estudios Turolenses.

Pacios Lozano, Ana R. (2002). *Bibliografía de arte mudéjar. Adenda 1992-2002*. Teruel: Centro de Estudios Mudéjares.

Pacios Lozano, Ana R. *Bibliografía de arte mudéjar. Adenda 2003-2013*. (unpublished)

Pacios Lozano, Ana R. y Carlos García Zorita (2016a). "Una aproximación histórico-bibliométrica a la producción científica del arte mudéjar (I)". *Quiroga: Revista de Patrimonio Iberoamericano* 9 (enero-junio 2016) :78-92. <http://revistaquiroga.andaluciayamerica.com/index.php/quiroga/issue/view/10>

Pacios Lozano, Ana R. y Carlos García Zorita (2016b). "Una aproximación histórico-bibliométrica a la producción científica del arte mudéjar (II)". *Quiroga: Revista de Patrimonio Iberoamericano* 10 (julio-diciembre 2016): 70-81. <http://revistaquiroga.andaluciayamerica.com/index.php/quiroga/issue/view/9>

Cita:

GARCÍA ZORITA, C. & PACIOS, A. R. (2018). Topic modelling characterisation of Mudejar art based on document titles. *Digital Scholarship in the Humanities*, Volume 33, Issue 3, pp. 529–539. <https://doi.org/10.1093/lhc/fqx055>

Robinson, L., E. Priego and D. Bawden (2015). “Library and information science and digital humanities: two disciplines, joint future?”. *14th International Symposium on Information Science*, 19-21 May, Zadar, Croatia. <http://openaccess.city.ac.uk/11889/>

Romero Frías, Esteban y María Sánchez González, eds., (2014). *Ciencias Sociales y Humanidades Digitales. Técnicas, herramientas y experiencias de e-Research e investigación en colaboración. CAC, Cuadernos Artesanos de Comunicación* 61. <http://www.cuadernosartesanos.org/2014/cac61.pdf>

Schreibman, Susan, Ray Siemens and Unsworth, John, eds. (2004). *A companion to Digital Humanities*. Oxford: Blackwell. <http://www.digitalhumanities.org/companion/>

Weingart, Scott (2017). *Topic Modeling for Humanists: A Guided Tour*. Web. 22 June <http://www.scottbot.net/HIAL/index.html@p=19113.html>

Wuchty, Stefan, Benjamin F. Jones and Brian Uzzi. (2007). “The Increasing Dominance of Teams in Production of Knowledge”. *Science* 316.5827: 1036–1039.

Jockers, Matthew L. and David Mimno (2012). *Significant Themes in 19th-Century Literature*. [preprint]. Lincoln: University of Nebraska, Faculty Publications, Department of English, August. <http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1105&context=englishfacpubs>

Valdés Fernández, Manuel (2006). *La arquitectura mudéjar y los sistemas constructivos en los reinos de León y Castilla en torno al 1200*. María Carmen Lacarra Ducay, coord. *Arte mudéjar en Aragón, León, Castilla, Extremadura y Andalucía*. Zaragoza: Institución Fernando el Católico. http://ifc.dpz.es/recursos/publicaciones/26/12/_ebook.pdf

Cita:

GARCÍA ZORITA, C. & PACIOS, A. R. (2018). Topic modelling characterisation of Mudéjar art based on document titles. *Digital Scholarship in the Humanities*, Volume 33, Issue 3, pp. 529–539. <https://doi.org/10.1093/lhc/fqx055>