Archeologia e Calcolatori
34.2, 2023, 9-32
doi 10.19282/ac.34.2.2023.01

DEVELOPING A DIGITAL ARCHAEOLOGY CLASSIFICATION SYSTEM USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TECHNIQUES

1. The scientific background

This article draws on an interdisciplinary research project promoted in 2019 to enhance the contribution of the National Research Council of Italy in the field of Cultural Heritage by means of a comparative study with the leading Italian and international research players. The simultaneous establishment of the Institute of Heritage Science (ISPC), which represents today the CNR's hub for research, innovation, and technological transfer of the Cultural Heritage strategic area, made it possible to launch an attempt to systematically classify its main expertise. The study was entrusted to SIRIS Academic, a consulting company specialised in higher education and research policies, which has been involved for several years in the development of research portfolio analyses – as a means of characterising research, based on the semantic content of scientific production and research projects. Since existing classification systems (e.g. bibliometric categories) do not represent contemporary research (mainly multidisciplinary, challenge-based), new transversal approaches that directly explore the semantic content (e.g. titles, abstracts, summaries) of research outputs have been developed by SIRIS Academic using text mining methods. To this end, the company has cooperated with numerous universities, research centres, governmental bodies and research quality assessment agencies (see Zenodo SIRIS: https://zenodo.org/communities/siris-academic/).

The analysis of a large corpus of textual data of the ISPC, extracted from research and project activities, was conducted in collaboration with an interdisciplinary team of ISPC researchers (Caravale *et al.* 2021). The study was divided into two distinct, closely interrelated levels: a) identifying the research competences on Heritage Science within the CNR (including the ISPC), and b) comparing them with the competences at the national level, considering other research organisations (e.g. INFN, INGV, etc.) and universities.

The research process involved a set of successive steps, which started with the identification and extraction of CNR research projects and publications between 2010 and 2019 inclusive, using a controlled vocabulary (hereafter, VOC) for Cultural Heritage. A controlled vocabulary is an organised arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching (Harping 2013). In other words, the goal of a VOC is to fully identify a specific research area/topic based on a given definition that can be adapted to a particular scope often difficult to find in traditional classification systems (e.g. Neuroscience, encompassing neurobiology, neurology, mental disorders, mental health/wellbeing and its social aspects). The specific perimeter of a VOC is established with field experts ahead of the analysis, to construct a conceptual map that defines the boundaries of a specific domain of interest. The presence of a term from the VOC (in title, abstract and/or author keywords), identifies the document as pertaining to the domain (Fuster, Massucci, Matusiak 2020).

The following were chosen as data sources: the Scopus citation database for publications; the CORDIS Community Research and Development Information Service for projects funded under the European Union's framework programmes (FP7 and H2020); and finally, the database of projects of the Creative Europe programme, all available in open format. As for ISPC, since the institute is significantly characterised by non-bibliometric research domains whose publications are often not indexed in international databases, information from PEOPLE, the CNR platform that hosts the institutional repository of research products, was added.

Data were also supplemented with information on the electronic resources stored in the repository of the CNR open access journal «Archeologia e Calcolatori» specialised in computational archaeology. During its 30-year publishing activity, the journal has classified articles using a dual taxonomy: the typology of computer applications to archaeology and the archaeological research fields largely involved in the application of computer methods. It could therefore serve as a well-established reference example of a scientific-academic classification implemented in a top-down approach by cross-domain experts and based on their knowledge of specific theoretical and methodological issues.

Of particular interest for its heuristic implications was the research phase addressed to content analysis and the identification of the most relevant topics dealt with in publications (topic modelling). Topic modelling is a machine learning technique that serves to automatically 'discover' the topics from a collection of texts. It is a bottom-up, automatic and unsupervised technique and is very useful for conducting an emerging analysis of research, technology and innovation ecosystems. This method applies to un- or semi-structured texts and makes it possible to identify on a statistical probabilistic basis the co-occurring lexical clusters (topics) that characterise a collection of documents and to analyse their distribution.

At this point, the intervention of experts becomes crucial: unsupervised machine learning methods are mainly used in the exploratory phase, when the aim is to extract from the data some otherwise not readily discernible structures and to highlight associations between topics sharing a common terminology but apparently unrelated. The re-classification and structuring of the results of this procedure, also in view of proposing interpretations and making predictions, is the task of the researchers, who should identify a model to guide and rule the investigation.

In examining the results, it was interesting to observe how the subject-specific topics recorded in the A&C classification were absorbed within the broader field of Cultural Heritage. The first factor that clearly emerged was that the specificity of individual research approaches featuring the field of digital archaeology is getting blurry, as records were largely merged into broader and more wide-ranging topics. For example, the marked preponderance of field research methods that emerges from the A&C classification resulted less evident, being distributed among different topics, from the more general 'Archaeological Research and Methods' to the more specific 'Geophysics, Digital Mapping and GIS', 'Photogrammetry, Image Processing and Digital Reconstruction' and 'Remote Sensing', which emphasise the technical-scientific aspect of the research though not including the more traditional closely humanistic one.

For this reason, it was decided to reconsider the analysis of the data extracted from the journal, by focusing precisely on the topic of digital archaeology with the specific aim of finding new ideas to supplement the classification which was first drafted over twenty years ago (Moscati 1999) and which has since followed the evolutionary course of the discipline (Cantone, Caravale 2019). This is the aim of the current study.

Digital archaeology contributes significantly to the more general scenario of Heritage Science, as evidenced by the laboratories of the Italian node of the E-RIHS infrastructure and by the activities recently promoted as part of the PNRR H2IOSC (Humanities and cultural Heritage Italian Open Science Cloud) project launched at the end of 2022. By stimulating the production of research perspectives that look to the past but are in line with the breakthrough development of science and technology (Moscati 2021), digital archaeology represents a specific research area, with a highly interdisciplinary character, supported by a long tradition of studies. At the same time, it forms an integral part of an innovative process of growth and development combining research, conservation, and scientific dissemination, in close relationship with the needs and requirements of today's information society.

2. Datasets

For the present analysis, we rely on two main datasets: the open access repository of «Archeologia e Calcolatori», which is registered in the list of OAI Data Providers (http://www.archcalc.cnr.it/oai/aec_oaipmh2.php) and the publications of the 'Computer Applications and Quantitative Methods in Archaeology' conference proceedings and journal. The first one is the focus of our analysis, whereas the latter is used as a benchmark for the former.

2.1 *A&C*

The data from publications in «Archeologia e Calcolatori» (A&C) – coming from both regular issues and Supplements – were provided directly by the editors of the journal and consist, most importantly, of their title, abstract, year of publication and affiliations of the authors. After some light data pre-processing (such as removal of records without much textual content, e.g. Introduction, Preface, etc.), and limiting ourselves to the period 2011-2020, the dataset counts with 477 records.

From a methodological point of view, the second decade of the new Millennium marks the consolidation of certain methods and the development of new ones. In the same period, from an editorial point of view, A&C witnessed the publication of as many as 9 conference proceedings (http://www.archcalc.cnr.it/pages/special_issues_proceedings.php) in addition to the regularly submitted articles. Furthermore, in 2019 a special issue was dedicated to the 30th anniversary of the journal (Moscati 2019). As for the Supplements, the data contains the 10 volumes published in the decade under investigation (http://www.archcalc.cnr.it/supplements/year_list_sup.php).

2.2 *CAA*

The publications in the 'Computer Applications and Quantitative Methods in Archaeology' conference proceedings and journal (CAA) constitute another long-standing observatory of key trends in computational archaeology, making it suitable as a benchmark for our study. The annual meetings originated in England in the 1970s and have grown over time, becoming today an international event open to large numbers of participants (Caravale, Moscati 2021, in part. 91-94). The high number of sessions at the conference that celebrated the CAA's 50th anniversary in Amsterdam in April 2023 is valuable evidence of its popularity today.

CAA too is actively pursuing the open and free access to all its proceedings volumes. Digital versions up to 2011 can be accessed via the online Proceedings portal (https://proceedings.caaconference.org/), and from 2012 to 2017 at the CAA Proceedings Bibliography web page (https://caa-international.org/publications/proceedings/bibliography/).

The CAA data was, for the most part, provided by the A&C editors, but was also supplemented by querying OpenAlex (Priem, Piwowar, Orr 2022). For present purposes the most relevant features of this dataset are the title, abstract and year of publication. After some light data pre-processing (such as removal of duplicates and nulls), and once again limiting the data to the period 2011-2020, the dataset consists of 514 records. Note that the records pertaining to the period 2011-2017 correspond to CAA conference proceedings, whereas those from 2018 until 2020 correspond to articles

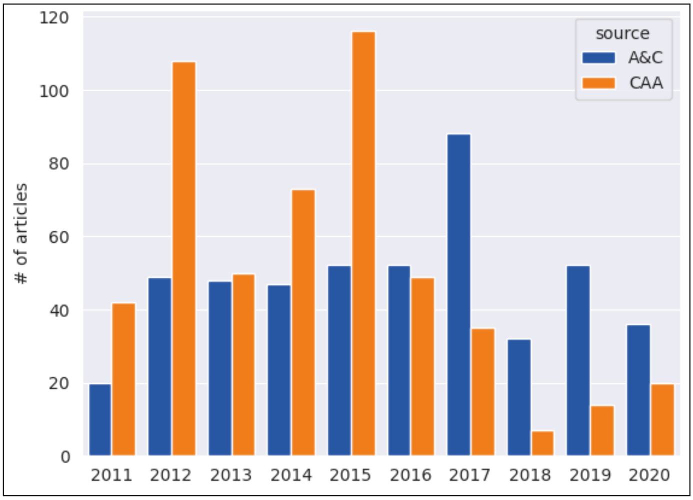


Fig. 1 – Bar chart showing the number of articles per year distinguished by source.

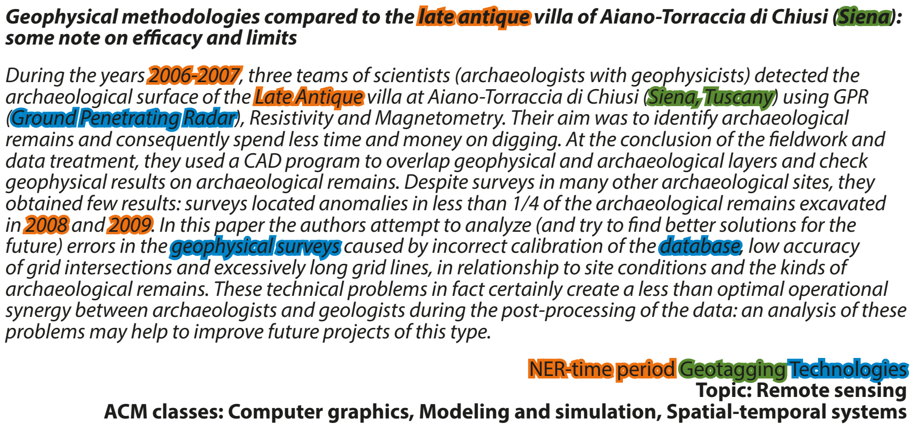
published in the newly launched «Journal of Computer Applications in Archaeology» (JCAA).

In Fig. 1, we see the number of articles per year distinguished by source. Numbers are relatively constant for A&C (with the exception of a significant increase in 2017, which is explained by the publication of two issues, the second one dedicated to the proceedings of the KAINUA 2017 International Conference on Knowledge, Analysis and Innovative Methods for the Study and the Dissemination of Ancient Urban Areas: Garagnani, Gaucci 2017). On the other hand, we see a bigger variation for CAA. In this case, the increase in 2012 and 2015 can be explained by the 40th anniversary of the conference (Earl *et al*. 2013) and the highly attended edition held in Siena (Campana, Scopigno, Carpentiero 2016), respectively. The drop in the last three years is due to the fact that the publication of the proceedings is paused since the 2018th edition, which is why the data for that period are extracted from the JCAA.

3. Methods

Our analysis of the A&C publications relies significantly on a prior process of information retrieval carried out by leveraging on multiple AI-based techniques. In this section, we describe the main methods employed and the results obtained, from a technical point of view.

The main tasks we have carried out, which will be described in detail shortly, are:


Geophysical methodologies compared to the late antique villa of Aiano-Torraccia di Chiusi (Siena): some note on efficacy and limits

During the years 2006-2007, three teams of scientists (archaeologists with geophysicists) detected the archaeological surface of the Late Antique villa at Aiano-Torraccia di Chiusi (Siena, Tuscany) using GPR (Ground Penetrating Radar), Resistivity and Magnetometry. Their aim was to identify archaeological remains and consequently spend less time and money on digging. At the conclusion of the fieldwork and data treatment, they used a CAD program to overlap geophysical and archaeological layers and check geophysical results on archaeological remains. Despite surveys in many other archaeological sites, they obtained few results: surveys located anomalies in less than 1/4 of the archaeological remains excavated in 2008 and 2009. In this paper the authors attempt to analyze (and try to find better solutions for the future) errors in the geophysical surveys caused by incorrect calibration of the database, low accuracy of grid intersections and excessively long grid lines, in relationship to site conditions and the kinds of archaeological remains. These technical problems in fact certainly create a less than optimal operational synergy between archaeologists and geologists during the post-processing of the data: an analysis of these problems may help to improve future projects of this type.

NER-time period Geotagging Technologies
Topic: Remote sensing
ACM classes: Computer graphics, Modeling and simulation, Spatial-temporal systems

Tab. 1 – Example of a title and abstract from the A&C collection with the results of applying all techniques to them, including NER, Geotagging, Technology identification, Topic modelling and Supervised classification.

1. A supervised classification into subfields of computer science, based on data labelled with the Association for Computing Machinery (ACM) taxonomy.
2. Topic modelling, in order to discover emergent topics from the title and abstracts of the publications.
3. Technology identification, to match articles with the technologies mentioned in them.
4. Named Entity Recognition (NER) to identify specific entities that are relevant from an archaeological point of view.
5. Geotagging, to link articles with the geographical locations they are about.

In Tab. 1 we illustrate all of these tasks with an example of an article's title and abstract with all the corresponding extracted information.

3.1 *ACM classification*

For this task, our aim is to classify the A&C publications according to their topic within the field of Computer Science. To this end, we make use of data from the Association for Computing Machinery (ACM) taxonomy in order to train a text classification model. The ACM taxonomy (https://dl.acm.org/ccs) is a hierarchical ontology of subfields of computer science, including up to six nested levels of categories per document, and a total number of 1961 categories of all levels. There are two main taxonomies proposed by the ACM: one dating back to 1998 and a second one proposed in 2012. Our data is labelled with the latest one.

The ACM dataset employed contains 258.103 publications (with their associated title and abstract). As one would expect, the vast majority of categories are irrelevant for our field of interest, archaeology, so we decided to select a limited number of categories that were applicable. This selection was based mostly following the advice of the A&C editors. We selected 17 categories belonging to different levels of the hierarchy and treated all of them on a par, turning the task into a (non-hierarchical) multi-label classification problem.

The selected categories are: 'Computer graphics', 'Computer vision', 'Data mining', 'Decision support systems', 'Digital libraries and archives', 'Document management and text processing', 'Education', 'History of computing', 'Human-centered computing', 'Information retrieval', 'Machine learning', 'Modeling and simulation', 'Multimedia information systems', 'Natural language processing', 'Probability and statistics', 'Spatial-temporal systems', 'World Wide Web'.

The original ACM dataset was mainly composed by computer science and engineering publications. For this reason, we were concerned that a system trained with this data would not be able to generalise well in a domain as specific as archaeology. In order to check and potentially mitigate this bias, we trained models on three different training sets:

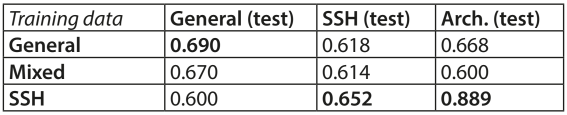
1) A general random sample of the original dataset.
2) A sample of publications which belong to Social Sciences and Humanities (SSH) domains. We did this by filtering the data according to whether they belonged to SSH related subcategories (e.g. 'anthropology', 'ethnography', 'economics', 'sociology', 'arts and humanities', 'fine arts').
3) A mix of the general random sample and the SSH publications.

Moreover, we split a test set for the general domain, one for the SSH domain and (a much smaller) one for the archaeology domain, simply filtering by the ACM category 'archaeology' (note that we removed the archaeology and the SSH-test data from the domain-specific and the general-domain training data).

For classification, as suggested in Cohan *et al.* 2020 and Singh *et al.* 2022, we train a linear support vector machine (SVM) on embeddings of concatenation of title and abstract, training a classifier for each category (Tab. 2).

Results show that in-domain training performs better both for general and SSH test data. In the case of publications in the archaeology domain, the classifier trained only on SSH publications performs much better than other options. Given the high F1[1] of this configuration, we take the model trained on the SSH dataset for our analysis of A&C and CAA publications.

[1] The F1 metric is widely used for evaluating classification tasks. It is defined as the harmonic mean of the precision and the recall metrics (where precision is the number of true positive results divided by the number of all positive results and recall is the number of true positive results divided by the number of all samples that should have been identified as positive). In other words, the F1 tells us both how correct and how complete our results are.



Training data	**General (test)**	**SSH (test)**	**Arch. (test)**
General	**0.690**	0.618	0.668
Mixed	0.670	0.614	0.600
SSH	0.600	**0.652**	**0.889**

Tab. 2 – Comparison of the macro-avg F1 between training data and test by domain.

3.2 *Topic modelling*

As briefly explained in the Introduction, topic modelling is an unsupervised machine learning technique that aims at automatically identifying texts that have semantic similarity and which is used to reduce complexity of textual corpora. Topic modelling techniques have been widely used for the identification of scientific topics in literature (Griffiths, Steyvers 2004; Callaghan, Minx, Forster 2020; García *et al.* 2020). From a technical point of view, while different methods and algorithms have been proposed to detect the topics, the use of pre-trained language models (PLM) based on Transformers such as BERT (Bidirectional Encoder Representation from Transformers: Devlin *et al.* 2019) is becoming increasingly popular for topic modelling (Stanik, Pietz, Maalej 2021; Sangaraju *et al.* 2022). In particular, we apply SPECTER (Cohan *et al.* 2020) – following the implementation in Bovenzi *et al.* 2022 – a BERT pre-trained model fine-tuned on scientific corpora and which also relies on citations in order to generate highly useful vectorial representations of scientific texts that produce embeddings for all texts (containing title and abstract), and then we use K-Means, an unsupervised clustering technique on top of the encoded vectors.

To find the best number of topics, we ran the K-Means by varying the number of clusters and we eventually chose to extract 10 clusters (i.e. topics) by qualitatively choosing the best trade-off between the semantic ‘richness’ of the topics and the overall number of topics (in order not to have neither too large clusters nor too little ones). Each cluster is therefore a topic and close vectors are thematically-related documents. Moreover, a domain expert manually selected a topic label for each topic. In Tab. 3, we list the final topics we detected together with the top keywords that appeared with most frequency in their documents. The lists of keywords have been slightly revised to provide a more intuitive description of the contents of the topics. We have deleted a few keywords that were irrelevant and that appeared across multiple topics.

3.3 *Technology identification*

An important type of information we retrieve from the titles and abstracts of the articles are the technologies mentioned in them. This level of analysis provides a new perspective with respect to the ACM classification in that we

	Topic label	Top keywords
0	**Artificial Intelligence**	neural, software, ontology, analytical archaeology, adaptive, computational, dataset, archaeology artificial, humanity
1	**GIS and spatial analysis**	archaeological datum, gis, geographical, archaeological information, geographic, archaeological research, documentation, archaeological site, software
2	**Imagery analysis**	image, reflectance, artefact, infrared, sense, multispectral, drawing, recognition, photograph, painting
3	**Material culture**	pottery, artefact, ceramic, archaeological site, archaeological datum, bone, stratigraphic
4	**Modeling and simulation**	simulate, settlement, computational, prehistoric population, climatic, gatherer, human foraging, geographical
5	**Digital cultural heritage**	cultural heritage, culture, archive, museum, archaeological datum, archaeological research, historical, digital cultural, humanities
6	**Photogrammetry and 3D scanning**	photogrammetry, photogrammetric, reconstruction, architectural, 3d model, scan, architecture, scanner, archaeological excavation, monument
7	**Remote sensing**	lidar, archaeological site, scan, geophysical, aerial, lidar datum, sense, remote sensing, airborne, terrain
8	**Semantic technologies**	archaeological datum, semantic, dataset, archive, archaeological information, documentation, semantic web, catalogue, software, archaeological database
9	**Virtual reality**	virtual, museum, interactive, immersive, reconstruction, vr, cultural heritage, archaeological site, artefact, exhibition

Tab. 3 – Most frequent keywords found in the documents belonging to each topic.

can go beyond computer science and extract insights about technology more generally. Moreover, it distinguishes itself both from the ACM classification and from the topic modelling in that the granularity achieved is a lot finer.

To perform this task, we make use of a controlled vocabulary (VOC). In particular, we use a VOC of technologies relevant for the field of Heritage Science (Duran-Silva *et al.* 2021), which was built by SIRIS Academic and the Istituto Regionale per la Programmazione Economica della Toscana, and which describes a set of key enabling technologies for culture and cultural heritage, many of which are in particular relevant for the field of computational archaeology (the list of key enabling technologies was mostly based on the proposal put forward in Borrione *et al.* 2019).

The list contains 905 keywords referring to technologies (such as 'machine learning', 'geographic information system', 'optical laser', '3D model', etc.), which are in turn classified in types (e.g. 'Lidar' is classified as belonging to the class '3D SCAN, PHOTOGRAMMETRY 3D/4D').

For each entry of the controlled vocabulary, we query the Wikipedia API (https://github.com/goldsmith/Wikipedia) and look for its corresponding suggested articles, the summaries of which we then vectorize (with SPECTER-2; Singh *et al.* 2022) to obtain their embeddings. These embeddings are compared with the embedding of the Wikipedia entry for 'Computational Archaeology' and the one with the highest cosine similarity with respect to it is chosen as the correct one, i.e. the one capturing the definition of that technology, if it is high enough in the Wikipedia results list[2]. This method allows us to disambiguate between possible meanings of the technologies' names.

To give an example, when we search for 'drone' in Wikipedia, we get the following candidate results:

Drone (bee) – a male honey bee.
Unmanned aerial vehicle – a generic drone.
Delivery drone – a drone used to transport packages.
Drone warfare – a form of aerial warfare using unmanned combat aerial vehicles.
Unmanned combat aerial vehicle – a combat drone.
Drone (sound) – a type of sound used in some forms of music.
Drone music – a music genre.
Drone art – a form of art produced with drones.
Droners – a French animated series.

By looking at the semantic similarity between the embeddings of their summaries and the embedding of 'Computational Archaeology' and taking also into account the position of the entries in the list of results, we obtain, correctly, that the relevant entry is B.

Once we have an embedding for each technology listed in our VOC, we proceed to match the technologies with the titles and abstracts of the publications. This involves a two-step process. First, we look for matches amongst the noun phrases contained in the titles and abstracts (these matches are fuzzy[3], in the sense that they allow for degrees: the higher they are, the less spelling differences between the terms). Second, we compute the cosine similarity between the SPECTER-2 embeddings of the titles and abstracts and the technologies' embeddings (obtained in the previous step). Finally, we use a manually set threshold depending both on the degree of fuzzy match and the embeddings' similarity in order to decide whether each noun phrase matches any of the technologies. Thus, whether an article is about a certain technology or other is decided both by looking at the degree of match of strings of characters, but also at the degree of semantic similarity between the meaning of the titles and abstracts and the descriptions of the technologies. This combined measure provides the right balance between looking only at purely surface morphological features (often too strict a criterion) and looking only into the semantic representation of concepts (often too imprecise a criterion).

[2] The reason why we take into account not only similarity with the meaning of 'Computational Archaeology', but also the position in which an entry appears in the list of results is that Wikipedia provides results sorted by relevance, which is often a useful measure for us, since completely irrelevant results are very unlikely to be the ones we are looking for.

[3] To do fuzzy matching, we use the Fuzzywuzzy library, which can be found here: https://github.com/seatgeek/fuzzywuzzy.

Entity	*F1-score*
Artefact	0.80
Time Period	0.81
Context	0.64
Material	0.68
Location	0.73
Specie	0.69
Overall	0.80

Tab. 4 – The F1-score obtained for each category of named entities.

3.4 *Archaeological NER*

Named Entity Recognition (NER) is the task of identifying important entities mentioned in an unstructured text. This type of information extraction allows the access to a finer level of granularity than techniques like supervised classification or topic modelling. Generic NER tasks focus on extracting expressions such as person names, locations, organisations, time expressions or quantities. However, for this study we trained a specialised NER model, which extracts only archaeologically relevant entities. To this end, we built on previous work in which an archaeological NER was trained on excavation reports in Dutch (Brandsen *et al.* 2020). Our strategy was to first automatically translate the Dutch training data into English, with the DeepL API (https://www.deepl.com/), and then retrain a NER model based on it.

The evaluation metrics obtained training on 80% of the datasets and evaluating on the other 20% are summarised in Tab. 4. Note that the results in English are more competitive than the results for Dutch text reported by the original paper (Brandsen *et al.* 2020).

For the purposes of our analysis, the type of identified entity that proved most relevant was Time Period, so in the Results and Discussion sections, we focus solely on it.

3.5 *Geotagging*

By geotagging a text we can get, in an automatic way, the geographical scope of a document (Andogah, Bouma, Nerbonne 2012), which can be especially interesting in a field like archaeology. More precisely, geotagging consists of 1) the identification of geographic entities in a text, and 2) toponym resolution, namely, linking them with their corresponding spatial location. The first part is a special case of NER. In order to perform this task, we make use of two pre-trained and openly available models: GeoText (https://github.com/elyase/geotext) and Geograpy3 (https://github.com/somnathrakshit/geograpy3) (we join their outputs to obtain more comprehensive results).

In order to perform toponym resolution we match the identified locations against the geographic database Geonames (via the Local-geocode library) (https://github.com/mar-muel/local-geocode), thereby obtaining their geographic coordinates. Note that, even though GeoText and Geograpy occasionally identify entities which are not locations, the process of toponym resolution filters out most of the previously introduced errors.

4. Results

4.1 *ACM classification*

The first set of results concerns the predictions of our ACM classifier. In Fig. 2 we can see the distribution of predicted categories on the A&C publications. The model predicts a high number of publications to be about Human-centered computing. This is explained from the fact that this is a level 0 category within the ACM taxonomy which encompasses many relevant sub-categories for us (interaction design, virtual reality, social media,

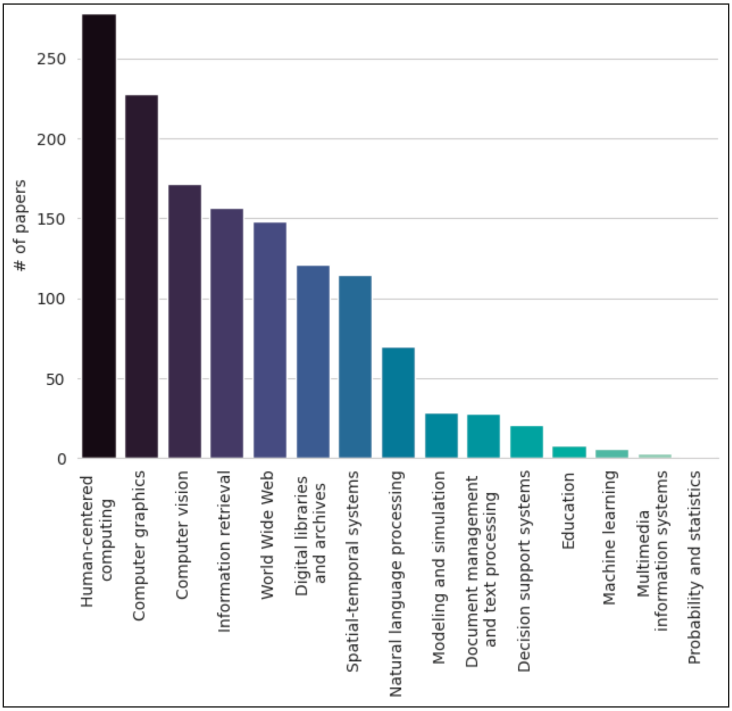


Fig. 2 – ACM classification: distribution of predicted categories on the A&C publications.

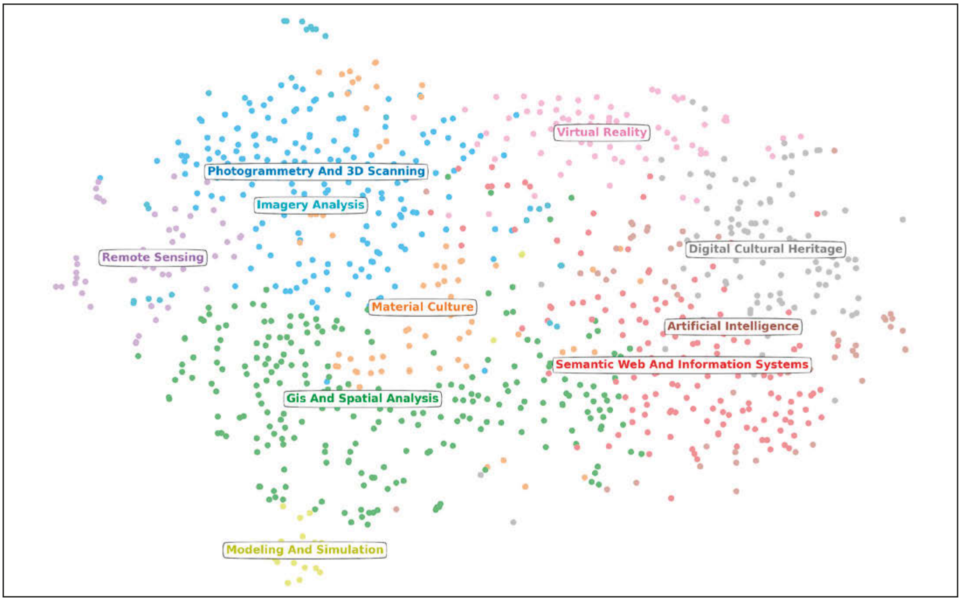


Fig. 3 – Emerging topics: results of the topic modelling as applied to the A&C and CAA data.

collaborative computing, visualisation, etc.). Another notable category is Computer vision. This category must be understood in a broad way, since it contains sub-categories such as Image representation, Image and video acquisition or 3D imaging, thus under the ACM definition it is highly instantiated in our dataset.

4.2 *Emerging topics*

In Fig. 3, we have mapped the results of the topic modelling as applied to the A&C and CAA data. It can be seen as a mapping of the field of digital archaeology from 2011 to 2020. In the plot, each dot represents a publication, the colours represent each of the 10 identified topics and the distance between dots represents semantic similarity between the titles and abstracts of the different publications. Consequently, the proximity between topics must also be interpreted as capturing their similarity.

In Fig. 4, we can see more clearly the amount of papers that fall under each topic and, moreover, we can see it separately for each publication. Unsurprisingly, GIS and Spatial Analysis is the most populated topic for both sources. Moreover, even though there are some disparities, A&C seems to be well represented in all identified topics, with the exception of Imagery analysis and Modeling and Simulation, where CAA has a significantly higher production.

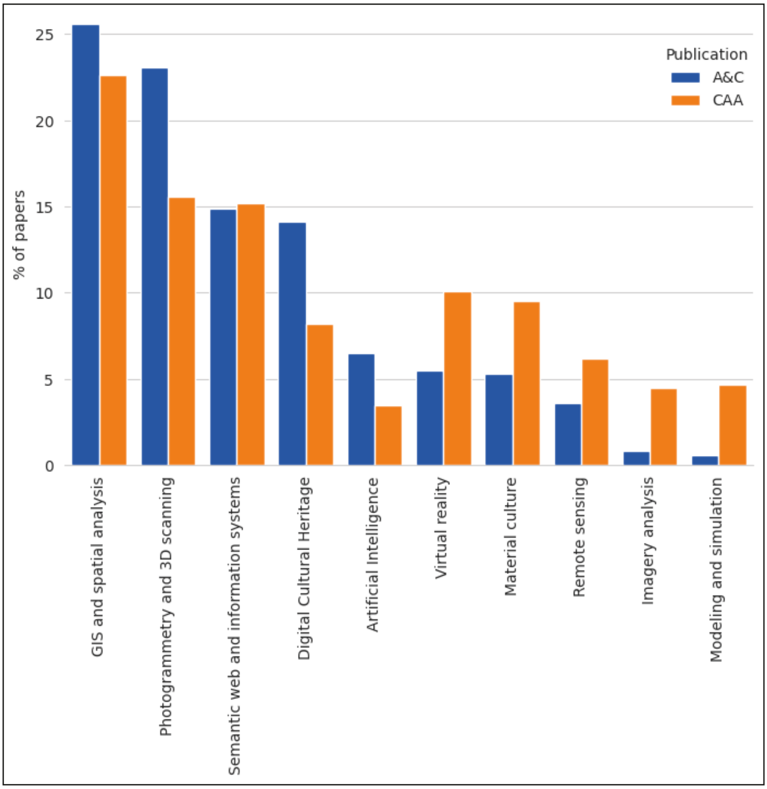


Fig. 4 – Amount of papers that fall under each topic, shown separately for each publication.

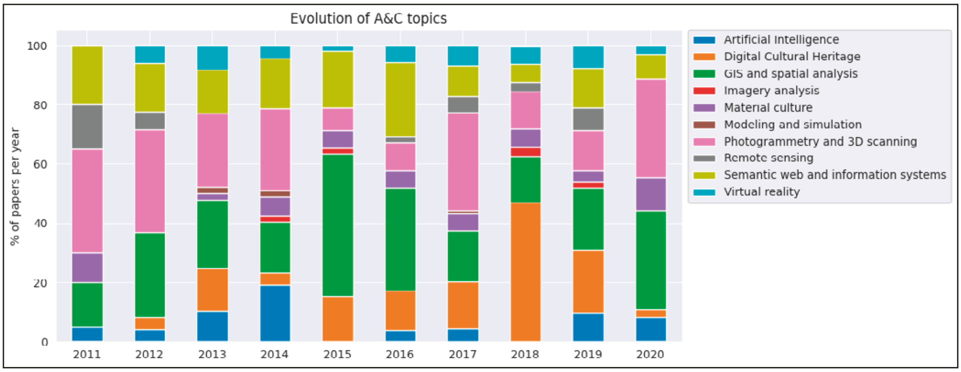


Fig. 5 – Chronological evolution of topics in A&C publications, represented as percentages of the number of articles per year.

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Artificial Intelligence	1	2	5	9	0	2	4	0	5	3
GIS and spatial analysis	3	14	11	8	25	18	15	5	11	12
Imagery analysis	0	0	0	1	1	0	0	1	1	0
Material culture	2	0	1	3	3	3	5	2	2	4
Modeling and simulation	0	0	1	1	0	0	1	0	0	0
Digital cultural heritage	0	2	7	2	8	7	14	15	11	1
Photogrammetry and 3D scanning	7	17	12	13	4	5	29	4	7	12
Remote sensing	3	3	0	0	0	1	5	1	4	0
Semantic web and information systems	4	8	7	8	10	13	9	2	7	3
Virtual reality	0	3	4	2	1	3	6	2	4	1

Tab. 5 – Time evolution of the absolute numbers of A&C publications belonging to each topic.

Finally, in Fig. 5 the time evolution of topics in A&C publications is represented (as percentages of the number of articles per year), while in Tab. 5 we can see the same data in absolute numbers represented in tabular form.

4.3 *Identification of technologies*

In Fig. 6, we answer two questions: what technologies are most frequently mentioned in A&C articles and how do their relative frequencies compare with their frequencies in CAA. In particular, we see represented the Top 20 most cited technologies in A&C. Finally, we can check the evolution in time of the technology mentions in A&C. In the following barplot (Fig. 7), we see the evolution of the 5 most mentioned technologies.

4.4 *Time periods prevalence*

By using NER, we were able to identify the time periods mentioned in A&C publications, bearing in mind that we have only taken into account terms that appear at least twice in our corpus. Moreover, we have grouped specific time periods into the 6 main groups that are represented in the figures, for ease of interpretation. For instance, 'Roman' was classified as CLASSICAL ANTIQUITY, 'Bronze Age' as PROTOHISTORY and 'Romanesque' as MIDDLE AGES (Fig. 8). If we do benchmarking with these results, we observe that CAA contains considerably less mentions of time periods across the whole classification (Fig. 9).

4.5 *Geographical scope*

Finally, Fig. 10 answers the question of what is the geographical scope of the A&C and the CAA publications. In this map, we project each location mentioned in A&C (red) and CAA (blue). We notice, as expected, that A&C contains a high concentration of locations in Italy. To see it more clearly,

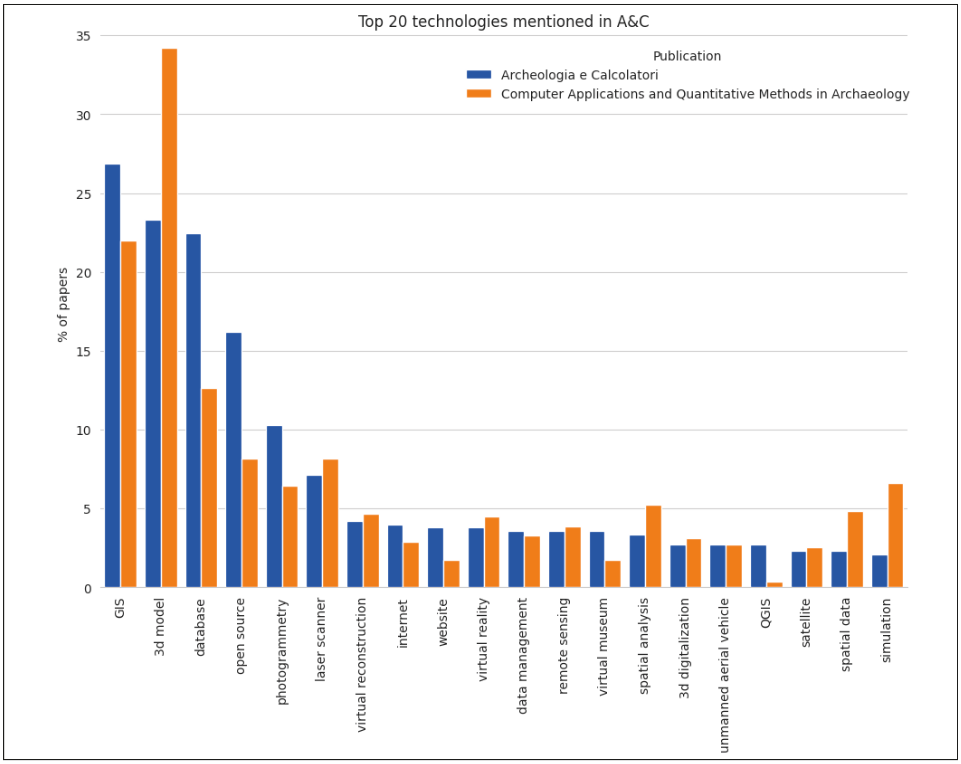


Fig. 6 – Barplot showing the Top 20 most cited technologies in A&C and their relative frequencies compared with those cited in CAA.

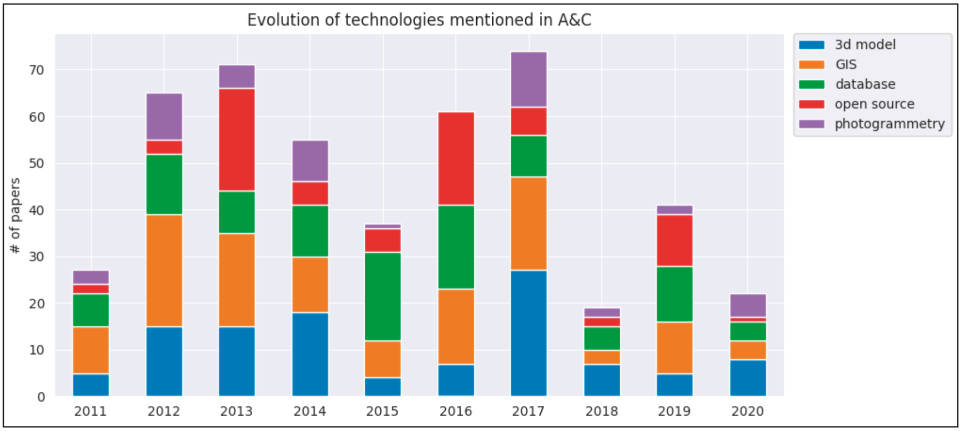


Fig. 7 – Barplot showing the evolution of the 5 most mentioned technologies in A&C.

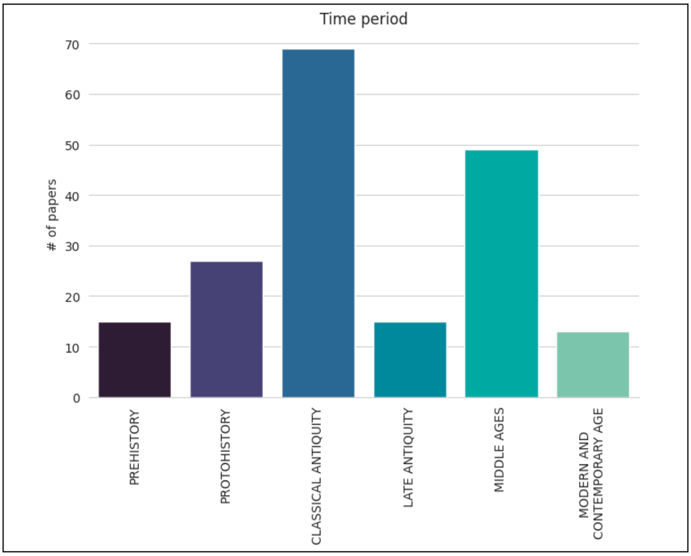


Fig. 8 – Numerical overview of the time periods mentioned in A&C publications.

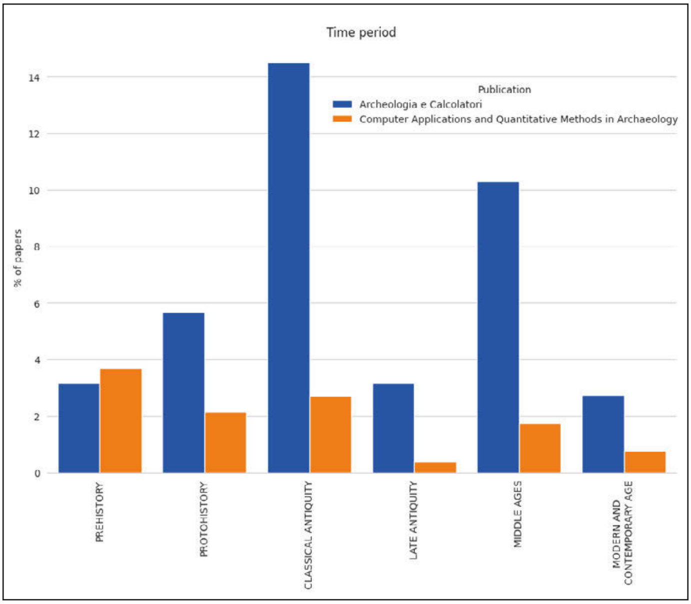


Fig. 9 – Amount of papers mentioning time periods in A&C and CAA data.

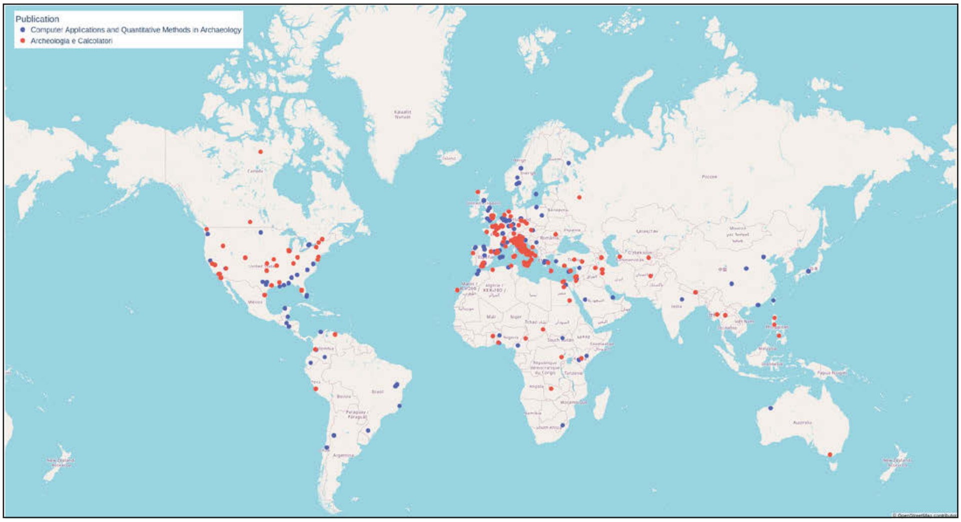

Publication
● Computer Applications and Quantitative Methods in Archaeology
● Archeologia e Calcolatori

Fig. 10 – Geographical scope of the A&C and the CAA publications. On the map, each location mentioned in A&C (in red) and CAA (in blue) is projected.

we zoom in on Italy (Fig. 11). There are a few CAA locations, but the vast majority correspond to A&C publications.

5. Discussion

The results illustrated in the previous section are a further step towards the new classification of the main cross-cutting themes featuring computer applications in archaeology in the Third Millennium that we have been dealing with since the publication of the A&C 30th issue. Indeed, the recent trend of digital archaeology to merge into the broader fields of Digital Cultural Heritage and Heritage Science implies a change of course as far as the description, distribution and classification of the application fields of computer science to archaeology are concerned. They are strongly informed by the rapid and compact progress of STEM disciplines on the one hand, and the Creative Industry on the other, resulting in a broad spectrum of technological innovations, which seem to escape any attempt at systematic classification.

The previous content analysis of the articles published in the journal and its Supplements over the last two decades was conducted using geographical text mapping strategies, multidimensional analysis techniques and the Social Network Analysis, to explore the relationship between archaeological themes and information technologies. The analysis illustrated in this paper focuses mainly on technological aspects and its first outcome allows us to check the

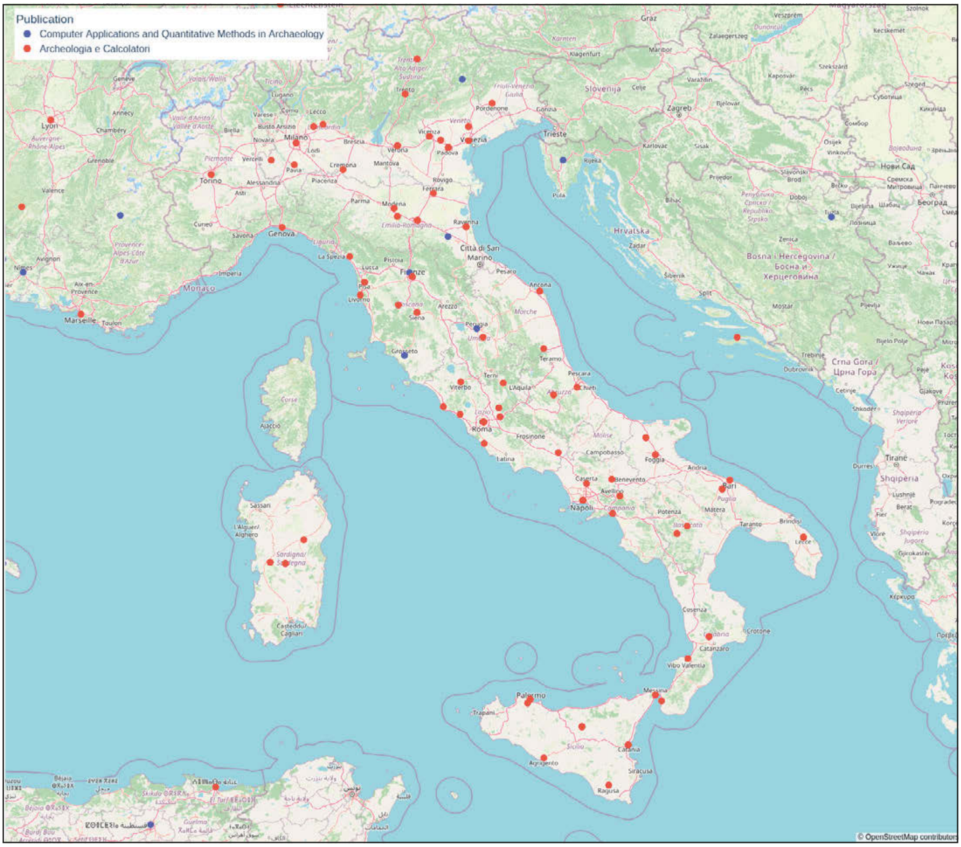

Publication
● Computer Applications and Quantitative Methods in Archaeology
● Archeologia e Calcolatori
© OpenStreetMap contributors

Fig. 11 – The same results as in Fig. 10, zoomed in on Italy.

consistency between the entries of the journal's list of 10 ICT topics[4] and what can be achieved by the application of various AI-based techniques, with both descriptive and predictive purposes.

Before discussing the results, a premise is necessary. As far as the comparison between A&C and CAA datasets is concerned, the results should take into account the publishing venue and its scope. A scholarly journal generally publishes the papers like they are spontaneously submitted by the authors, and thus, by its very nature, it tracks ongoing methodological trends and their evolution. In contrast, conference proceedings are often grouped according

[4] Computer graphics IP CAD, Data encoding and metadata, Database, GIS and cartography, History of applications and research projects, Multimedia and web tools, Remote Sensing, Simulation AI, Statistics, Virtual Reality and 3D modelling.

to the topics suggested by the organizers and, in recent years, by the 'call for sessions' process. In our case, the gap can be bridged to some extent thanks to the publication of conference proceedings and thematic issues within the journal and its Supplements, which provide grounds for comparison.

The first set of results, i.e. the supervised classification into subfields of computer science based on data labelled with the ACM taxonomy (Fig. 2), shows that the 15 selected categories have strong similarities with the A&C 10 ICT topics. Indeed, 'human-centered computing' corresponds to our history of applications and research projects, 'information retrieval' bears a strong relationship with database management systems, 'spatio-temporal systems' correspond to GIS, 'computer graphics' and 'computer vision' find a strong correlation with CAD, BIM and Virtual Reality applications. Not surprisingly, applications related to AI (modelling and simulation, decision support systems, machine learning techniques) continue to be statistically less frequent in archaeology, while it is worth analyzing the expected future decline of methods labelled 'probability and statistics'. These methods, which underlie the rise of quantitative archaeology in the 1960s, are currently part of the research practice, and thus their application is often taken for granted. In the last issues of the journal, however, the presence of an increasing number of articles based on statistical techniques and their heuristic potential (more than 15 in 5 years) – maybe also due to the widespread use of the powerful and versatile R open source software – seems to contradict this result.

The outcomes of topic modelling are self-evident, but additional insights into the 10 clusters (i.e. topics) and the relative top keywords may be useful to suggest how to expand the classification of the journal's articles in the future. Let us take the topic 'Imagery analysis' as an example. Beyond the technical aspects, the list of keywords clearly identifies specific areas of advanced knowledge management, where interdisciplinary applications are closely intertwined with the development of AI tools. The scientific implications are manifold and impressive, ranging from the analysis and interpretation of remote sensing imagery to the automatic recognition of archaeological potsherds.

Data on the evolution of the 5 most frequently mentioned technologies in A&C (Fig. 8) can be read as indicators of two well-defined issues in today's digital archaeology: the use of more 'traditional' and well-established methodologies, such as GIS and databases – which over time have maintained their role as key tools in the management of archaeological data – alongside more recent and progressively emerging techniques, such as digital photogrammetry and three-dimensional modelling. Their frequent occurrence in the recent issues of the journal, partially interrelated with the publication of conference proceedings specifically dedicated to these topics, underlines their success as tools for documenting the past, at a time when visual data have acquired a key role in both the research and the dissemination of archaeological results.

For instance, ISPC, which coordinates the Italian node of the European research infrastructure on Heritage Science E-RIHS.it, has many research Labs dedicated to these topics (https://www.ispc.cnr.it/en/ricerca/gruppi_e_labs/).

The case of the 'open source' entry is quite different. With respect to software development, this undoubtedly indicates its cross-cutting role in interacting with the other four technologies. Its presence can be associated with the journal's close cooperation with the ArcheoFOSS community, which since 2006 has been promoting open tools and technologies in the academic, professional and institutional Cultural Heritage domain, choosing the journal as a publishing venue for its workshops proceedings. However this result could also be linked to the dissemination of the wider open access movement, whose principles have been embraced by the journal since 2005 and today constitute the very lifeblood of the 'Open Data, Open Knowledge, Open Science' ISPC research Lab (https://www.ispc.cnr.it/en/2020/05/14/gruppo-open-data/).

Lastly, the results on geographical and chronological scope should deserve further investigation. The geographical distribution of the sites mentioned in A&C in the last five years (2014-2018) was investigated in 2019 thanks to the use of the Recogito tool as part of the Pelagios Network (Cantone, Caravale 2019). It can now be compared to data resulting from the automatic mapping (Figs. 10-11) to cover a large number of published issues. The territorial scope resulting from both analyses is wide, with a distribution throughout the Mediterranean area and beyond. A dominant role is no doubt played by the sites of our Peninsula, but a distinctive feature is definitely the great involvement of European Mediterranean countries, an achievement that vindicates the journal's pioneering choice to adopt the multilingualism approach. This finding emphasises the international nature of A&C, which focuses on wide-ranging initiatives in digital archaeology, in line with the policy traced in the opening editorial of the first issue, which stressed the need to exploit ongoing projects both in Italy and abroad (Cristofani, Francovich 1990).

Benchmarking with the results for the time periods mentioned in both A&C and CAA is a complex task, because the classification, as pointed out above, is currently too broad and will need to be further articulated. As for A&C, the prevalence of Classical Antiquity and Middle Ages fully matches the aims of the journal, which was launched in 1990 to collect and illustrate research projects conducted predominantly in the field of classical and post-classical archaeology. In any case, the chronological span is quite comprehensive, from prehistory to the modern age, indicating how widespread is the use of digital technologies in all fields dedicated to the study of Antiquity.

The reason that CAA publications contain only a few mentions of time periods across the whole classification is mainly due to the different publication venue. A&C, in fact, dedicates extensive articles (around 6000 words) to specific archaeological sites and monuments, comprehensively described in their geographical and chronological context, and, by its mission, only accepts papers giving equal emphasis to the archaeological and the technical aspects, in order to highlight the contribution of information technology to the development of archaeological research methods.

6. Conclusion

Although the results presented here can only be taken as a first approximation to be expanded in the future including the unpublished CAA conference Books of Abstracts and the last 4 issues of A&C, until reaching our ultimate goal, i.e. the analysis of the full texts of the A&C articles, they provide interesting insights into the latest advances in the field of digital archaeology, as discussed in the previous section.

This experiment shows the importance of combining different methods to better represent and analyse scientific literature, its contribution and evolution. The increasing availability of open metadata and open research information (eg. OpenAire or OpenAlex) offers a great opportunity to build more customised and multidimensional analyses beyond applying a predefined classification scheme. This semantic-based approach, including bottom-up information extraction techniques (such as Topic Modelling or Named-Entity Recognition) is particularly relevant for mapping and understanding the contribution of humanities and social sciences, such as archaeology (and especially so for interdisciplinary and rapidly changing fields, such as digital archaeology), where the use of traditional classifications (e.g. by discipline or journal) has demonstrated to be limited.

Alessandra Caravale, Paola Moscati
Istituto di Scienze del Patrimonio Culturale - CNR
alessandra.caravale@cnr.it
paola.moscati@cnr.it

Nicolau Duran-Silva, Berta Grimau, Bernardo Rondelli
SIRIS Lab, Research Division of SIRIS Academic
nicolau.duransilva@sirisacademic.com
berta.grimau@sirisacademic.com
bernardo.rondelli@sirisacademic.com

Acknowledgements

This work was supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya, by the EU HORIZON SciLake (Grant Agreement 101058573), and by the PNRR Project Humanities and cultural Heritage Italian Open Science Cloud - H2IOSC, CUP B63C22000730005. The ACM Digital Library dataset is provided for non-commercial research purposes, courtesy of the Association for Computing Machinery, Inc. (2023).

REFERENCES

Andogah G., Bouma G., Nerbonne J. 2012, *Every document has a geographical scope*, «Data & Knowledge Engineering», 81-82, 1-20 (https://doi.org/10.1016/j.datak.2012.07.002).

Borrione P., Carrozzino M., d'Orsi P., Iommi S., Lunghi M., Siotto E. 2019, *IRPET. Report della piattaforma "Tecnologie, Beni Culturali e Cultura". Le roadmap dello sviluppo e dell'innovazione (RIS3)*, Firenze, Regione Toscana (http://www.irpet.it/archives/53165).

Bovenzi N., Duran-Silva N., Massucci F.A., Multari F., Pujol-Llatse J. 2022, *Mapping STI ecosystems via open data: Overcoming the limitations of conflicting taxonomies. A case study for climate change research in Denmark*, in *Linking Theory and Practice of Digital Libraries, 26th International Conference on Theory and Practice of Digital Libraries, TPDL (Padua 2022), Proceedings*, Springer, Cham, 495-499 (https://doi.org/10.1007/978-3-031-16802-4).

Brandsen A., Verberne S., Wansleeben M., Lambers K. 2020, *Creating a dataset for Named Entity Recognition in the archaeology domain*, in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, European Language Resources Association, 4573-4577.

Campana S., Scopigno R., Carpentiero G. (eds.) 2016, *CAA2015. Keep The Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, Oxford, Archaeopress.

Cantone F., Caravale A. 2019, *Archeologia e Calcolatori. Classificazione geografica e tematica per la condivisione della conoscenza*, in Moscati 2019, 93-107 (https://doi.org/10.19282/ac.30.2019.07).

Callaghan M.W., Minx J.C., Forster P.M. 2020, *A topography of climate change research*, «Nature Climate Change», 10, 118-123.

Caravale A., Miliani C., Moscati P., Sfameni C. 2021, *La sfida delle competenze per il Patrimonio Culturale: complementarità, integrazione, interazione*, in A. Filippetti (ed.), *Le Scienze Umane, Sociali e del Patrimonio Culturale nell'era delle grandi transizioni*, Roma, CNR Edizioni, 65-86.

Cohan A., Feldman S., Beltagy I., Downey D., Weld D. 2020, *SPECTER: Document-level representation learning using citation-informed transformers*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2270-2282 (https://doi.org/10.18653/v1/2020.acl-main.207).

Devlin J., Chang M., Lee K., Toutanova K. 2019, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019, Minneapolis 2019)*, Association for Computational Linguistics, 1, 4171-4186 (https://doi.org/10.18653/v1/N19-1423).

Duran-Silva N., Parra-Rojas C., Rondelli B., Giocoli L., Plazas A., Grimau B. 2021, *A controlled vocabulary for research and innovation in the field of Cultural Heritage & Heritage Sciences [Data set]*, Zenodo (https://doi.org/10.5281/zenodo.8159015).

Earl G., Sly T., Chrysanthi A., Murrieta-Flores P., Papadopoulos C., Romanowska I., Wheatley D. (eds.) 2013, *Archaeology in the Digital Era: Papers from the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (Southampton 2012)*, Amsterdam, Amsterdam University Press (https://doi.org/10.2307/j.ctt6wp7kg).

Fuster E., Massucci F., Matusiak M. 2020, *Identifying specialisation domains beyond taxonomies: Mapping scientific and technological domains of specialisation via semantic analyses*, in R. Capello, A. Kleibrink, M. Matusiak (eds.), *Quantitative Methods for Place-Based Innovation Policy*, Cheltenham, Edward Elgar Publishing, 195-234.

Garagnani S., Gaucci A. (eds.) 2017, *Knowledge, Analysis and Innovative Methods for the Study and the Dissemination of Ancient Urban Areas, Proceedings of the KAINUA*

2017 International Conference in Honour of Professor Giuseppe Sassatelli's 70th Birthday (Bologna 2017), «Archeologia e Calcolatori», 28, 2 (https://doi.org/10.19282/AC.28.2.2017).

García D., Massucci F.A., Mosca A., Ràfols I., Rodríguez A., Vassena R. 2020, *Mapping research in assisted reproduction worldwide*, «Reproductive biomedicine online», 40, 71-81.

Griffiths T., Steyvers M. 2004, *Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America*, 101, Suppl. 1, 5228-5235 (https://doi.org/10.1073/pnas.0307752101).

Harping P. 2013, *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*, Los Angeles, Getty Publications.

Levy P. 1994, *L'intelligence collective. Pour une anthropologie du cyberespace*, Paris, La Découverte.

Moscati P. 1999, *"Archeologia e Calcolatori": dieci anni di contributi all'informatica archeologica*, «Archeologia e Calcolatori», 10, 343-352 (http://www.archcalc.cnr.it/indice/PDF10/10_23_Moscati.pdf).

Moscati P. 2019 (ed.), *30 anni di Archeologia e Calcolatori. Tra memoria e progettualità*, «Archeologia e Calcolatori», 30, 9-138 (http://www.archcalc.cnr.it/journal/idyear.php?IDyear=2019-01-01).

Moscati P. 2021, *Digital archaeology: From interdisciplinarity to the 'fusion' of core competences. Towards the consolidation of new research areas*, «Magazén», 2, 2, 253-274 (https://doi.org/10.30687/mag/2724-3923/2021/04/004).

Priem J., Piwowar H., Orr R. 2022, *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*, arXiv:2205.01833v2.

Sangaraju V.R., Bolla B.K., Nayak D.K., J. Kh 2022, *Topic modelling on consumer financial protection bureau data: An approach using BERT based embeddings*, in *2022 IEEE 7th International Conference for Convergence in Technology-I2CT (Mumbai 2022)*, IEEE, 1-6 (https://doi.org/10.1109/I2CT54291.2022.9824873).

Singh A., D'Arcy M., Cohan A., Downey D., Feldman S. 2022, *SciRepEval: A multi-format benchmark for scientific document representations*, arXiv:2211.13308v2.

Stanik C., Pietz T., Maalej W. 2021, *Unsupervised topic discovery in user comments*, in *2021 IEEE 29th International Requirements Engineering Conference (RE)*, IEEE, 150-161.

ABSTRACT

The Authors propose a knowledge map to analyse and access scientific contents related to Digital Archeology by leveraging various Machine Learning (ML) techniques. The case study concerns the articles published in our international journal «Archeologia e Calcolatori» in the decade from 2011 to 2020 and, as a benchmark, the publications in the 'Computer Applications and Quantitative Methods in Archaeology' (CAA) conference proceedings and journal. The titles and abstracts of the publications featured in these two data sets were analysed using a supervised classification approach into the subfields of computer science, based on the ACM's taxonomy, and by applying topic modelling techniques to discover emergent topics, Named Entity Recognition to identify specific archaeologically relevant entities, and geotagging techniques to link articles with the geographical locations they discuss. The results achieved, although preliminary, provide some methodological suggestions: i) the opportunity to build custom analyses by taking advantage of the increasing availability of open data and metadata; ii) the scope of the contribution of archaeology, and in particular of computational archaeology, to the Heritage Science interdisciplinary domain; the heuristic and predictive role of different ML techniques to gain a multi-faceted access to data analysis and interpretation.