

Off the record: Using data mining to review decision making in conservation practice

STAVROULA GOLFOMITSOU*

University College London Qatar
Doha, Qatar
s.golfomitsou@ucl.ac.uk

FLAVIA RAVAIOLI

University College London Qatar
Doha, Qatar
flavia.ravaioli.10@ucl.ac.uk

CATHERINE TULLY

University College London Qatar
Doha, Qatar
catherine.a.tully@gmail.com

GRAEME MCARTHUR

Wallace Collection
London, United Kingdom
mcarthur.graeme@googlemail.com

KATY LITHGOW

National Trust
Swindon, United Kingdom
Katy.Lithgow@nationaltrust.org.uk

*Author for correspondence

KEYWORDS: documentation, data mining,
museum databases, National Trust

ABSTRACT

Conservation records span years of past as well as current practice and can be used to inform future decision making in heritage organisations. This paper discusses the use of mixed quantitative and qualitative analytical techniques in querying digitised versions of such datasets. A subset of treatment records from the National Trust were used as part of a larger study investigating the rationale behind cleaning practices in museums and historic houses. Visualisation tools like Tableau Public and statistical analysis were employed to reveal patterns in the recorded treatments. By discussing the issues encountered in collecting, structuring and analysing such data, the authors address broader questions of standardisation, common terminology and information sharing in conservation.

INTRODUCTION

Records maintained by museums and other heritage institutions contain a wealth of information related to conservation practice, largely consisting of condition assessments, treatment reports and environmental data. The amount of information available has hugely increased since the 1980s as a result of the adoption of digital databases in museums and the gradual digitisation of past records, which has been supported by several international initiatives (Roy et al. 2007). Most condition and treatment information is in the form of text or numbers, entered into pro forma tables or as free text. Although a standardised structure is often missing in historic records, the information stored within has the potential to answer a broad range of questions. A variety of commercial and custom-made database software allows information to be input into a centralised, structured repository, thus increasing accessibility. The number of open-access software available is a testament to the growing demand for standardisation in heritage organisations (i.e. the 'Spectrum' standard for museum collections management published by the Collections Trust, <http://www.collectionstrust.org.uk/spectrum>).

The aim of this research is to develop a methodology for analysing conservation treatment records in the digital databases of heritage institutions. The study is part of University College London Qatar's (UCL Qatar) Coming Clean project, a four-year investigation into decision-making processes in conservation practice and the factors that affect them. More specifically, querying digital records will allow a better understanding of how different cleaning approaches are informed by particulars of practitioner context. A sample of digital records from the UK conservation charity The National Trust served as a case study to test the process of data retrieval, organisation and analysis.

Statistical methods have found wide applications in heritage and archaeology in the past half century (Hodder and Orton 1976, Reedy and Reedy 1988, Shennan 1997). In the conservation field, chemometrics in experimental design and sampling strategies has been employed (Golfomitsou and Merkel 2004, Sawdy and Price 2005a and 2005b). In addition, epidemiological approaches to collections management questions have been proposed by Suenson-Taylor and co-workers (1999), and have more recently become the focus of a project at the Getty Conservation Institute (http://www.getty.edu/conservation/our_projects/education/managing/epidemiology.html). Sampling

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

methodologies used to attribute cause to effect have been employed in the investigation and prioritisation of risks by English Heritage (Xavier Rowe et al. 2008), and statistics inform risk management approaches to collections management such as the CCI-ICCROM-RCE method (Waller 2003, Karsten et al. 2012). However, all of these depend on the use of data collected for the purpose of the analysis, rather than attempting to gain insights from historic ‘proxy’ data. The research reported here departs from previous approaches by shifting the focus of analysis to conservation practices and the drivers of decision making in heritage organisations.

THE NATIONAL TRUST’S COLLECTION MANAGEMENT SYSTEM

The National Trust has employed staff and freelance conservators to carry out condition reports on its collection since the 1970s, in addition to the observations made by trained property staff who are responsible for day-to-day monitoring and routine cleaning (Lithgow and Lloyd 2017). A bespoke database – the Collections Management System (CMS) – went live in 2009, specially commissioned to cope with the geographically dispersed nature of the organisation (over one million inventoried objects and decorative features in over 300 properties open to visitors). Having dealt with the first priority – the migration and cleaning of the inventory records on which conservation records depend – there is a large backlog of pre-2009 conservation records to be uploaded. A two-year pilot project completed in 2013 aimed to upload all the records in the Midlands region, which represent 13% of the Trust’s entire inventoried collection. The work was carried out by trained freelance conservators overseen by a National Trust member of staff with expertise in the CMS. Since 2015, a trained collections care member of staff was appointed to upload all completed records as well as providing pre-populated templates for condition and treatment records, again overseen by the CMS expert.

The CMS holds core conservation information within the object record on a specific tab (Figure 1), while more detailed or specialist reports can be uploaded to the same page as PDF/A format (the archive standard). Core


Inventory Number:	290226	Other Number:	CAL/P/43				
Area of Responsibility:	Calke Abbey, Derbyshire (Accredited Museum)	Description:	Oil painting on panel, Hounds putting...				
Object Category:	Art / Oil paintings	Creator:	Abraham Hondius (Rotterdam c.1625 – London 1691)				
Location:	10182.CAL: Midlands / Calke Abbey, Derbyshire / Calke Abbey /	Title:	Hounds putting up a Swan				
<div> Key Data 1 Key Data 2 Creation/Titles Associations Physical History Locations Conservation Media </div>							
Conservation History							
General Condition Note: + Slightly dusty. Crazed poorly saturating yellow varnish. Old retouched split at top left hand corner.		General Condition: B - Good General Stability: i - Stable General Priority: 2 - Treatment Desirable Monitoring Frequency: <not-set>					
Requirements Note: + <div></div>		Date: 01/06/2005 Name: Davis, Helen					
Return Displaying all 4 Reports < > x							
Report Id	Report Type	Date	Author	Reason	Status	Note	
91318	Condition	01/06/2005	Davis, Helen	Conservator Survey	Condition assessment complete	Slightly dusty. Crazed poorly saturating yellow varnish. Old retouched split at top left hand corner.	Show
41509	Condition	21/12/2004	Fahy, Jennifer			cleaned	Show
2959	Treatment	06/06/1990	Staniforth, Sarah	Backlog	Treatment completed		Show
119151	Prior-Condition	18/05/1985	Staniforth, Sarah	Conservator Survey	Condition assessment complete	Please see attached report for more detailed description of condition	Show
Return < > x							

Figure 1. Example of a conservation record in the National Trust’s CMS database

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

conservation information is common to every materials typology (e.g. description of condition, treatment proposals, stability and priority codes, hours and cost of treatment, author and date), and is partially structured into several fields, some consisting of drop-down menus, automated selection of 'fields' and the input of free text. Detailed reports are unstructured and presented in a variety of formats that are being converted to PDF/A over time. The first challenge encountered was therefore that of extracting information from proxy data that was in a variety of formats and identifying the limits this posed to analysis.

METHODOLOGY

The simple, though time-consuming solution to the variability encountered in the dataset was to manually select and copy the relevant information from the CMS forms as well as from attached documents. Four thousand and sixty condition and treatment reports referring to cleaning or removal of unwanted substances were selected from the Midlands' records. Examples of the keywords used are 'cleaned', 'dusted', 'removed', 'swabbed', 'dust' and 'dirt'. Excel was chosen for data storage because it is widely available, easy to use with semi-structured data and compatible with a variety of statistical software. Of the collected records, 1,625 treatment reports were manually coded according to predetermined categories to standardise and simplify key information for analysis. The coding system used had been developed by the authors for a survey of conservation professional literature on cleaning, the analysis of which is another outcome of this project. This will allow future comparison between cleaning techniques published in conservation literature and those used by conservators in practice. Coding consisted in reading through the text and selecting one of multiple options in Excel for categories such as 'Type of Object', 'Material', 'Substance Removed through Cleaning', 'Cleaning Method Employed'.

In the initial stage of analysis we kept an open-ended approach, employing exploratory techniques to identify broad trends and irregularities. This allowed formulating specific questions for further confirmatory analyses. The order in which the two types of software were used reflects this strategy. Visualisation graphs created with Tableau Public (<https://public.tableau.com/s/>) allowed initial observations to be made, which were followed up with statistical analysis using IBM SPSS Statistics (<https://www.ibm.com/us-en/marketplace/spss-statistics>). Tableau Public is the free available version of an interactive data visualisation tool which can transform complex data into intuitive pictures. Once shared online, interactive graphs allow the data to be explored by a wide audience through the use of filters and search terms. The interactive graphs created for this study are publicly available at www.comingcleanucl.com. Another example of the use of Tableau Public with interactive graphs can be seen at ICCROM's project Heritage Science (<http://www.iccrom.org/science-for-heritage/>). IBM SPSS Statistics is a widely used software for analysis in the social sciences. It was employed to identify statistical correlations and trends, attaching numerical values to associations between categories.

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

RESULTS OF DATA MINING

Graphs created with Tableau Public showed a clear pattern in the CMS records regarding the number of objects cleaned (Figure 2). While cleaning was primarily carried out on paintings up to the late 1970s, decorative art was treated more frequently after that, and from the mid-1980s cleaning occurred on a much greater variety of objects. Decorative art includes objects such as textiles, metals and ceramics, which are valued both for their aesthetic and functional properties. The observed timeline coincides with the development of a professional conservation service in the Trust, beginning with the appointment of a paintings conservator, who gradually appointed freelance advisers in other disciplines. A much larger number of objects are treated after 1999, a time in which the Trust increased its investment in collections conservation after reviewing the state of conservation of its assets in 1998–99 (Lithgow et al. 2008). The spike in social history objects treated in 2007 was due to a major project conserving objects required for the Sudbury Hall Museum of Childhood (Derbyshire).

The main substances removed are accidentally acquired surface deposits such as dust and other pollutants, both adhered and loose (Figures 3 and 4). Other common surface cleaning interventions involved the removal of past restorations, deteriorated varnish, accretions or corrosion. Biological deposits such as mould and bird droppings represent a small part of the cleaning treatments carried out. The bar chart in Figure 4 demonstrates which materials have been given most attention. This is a consequence

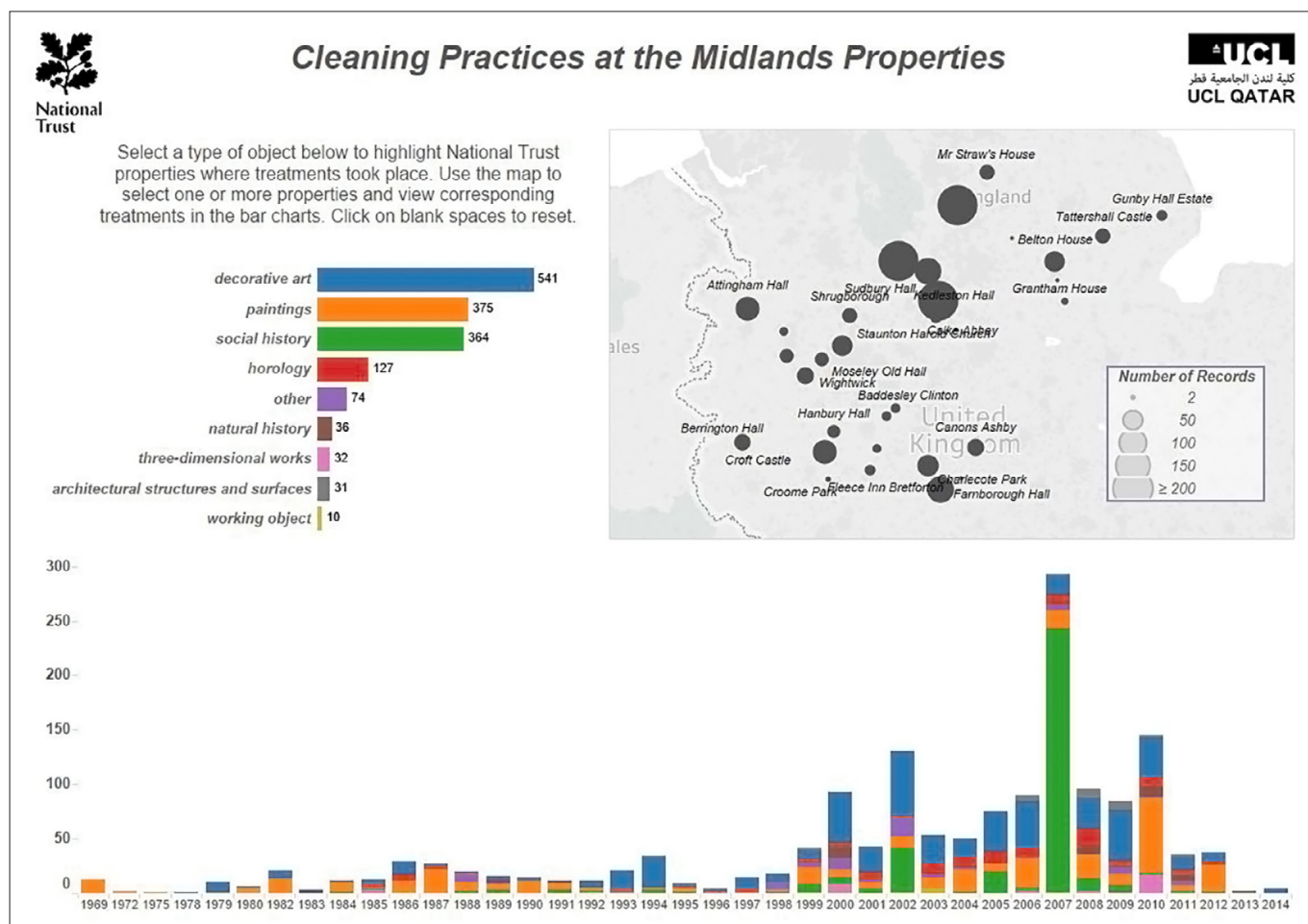


Figure 2. Rates of work illustrated over time across the Midlands properties, per type of object

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

both of the number of these types of material on open display, and of the value attached to them (paintings may not be as numerous as paper objects, but are prioritised for treatment). Cleaning techniques are illustrated by Figure 4, which shows the predominance of dry cleaning until year 2000, when more resources were invested in conservation. After that date there is greater evidence of wet cleaning and mixed methods, suggesting more complex treatments being undertaken.

IBM SPSS Statistics was used to further explore correlations between variables such as 'Type of Object', 'Material' and 'Cleaning Method Employed'. Of particular interest were possible patterns emerging from the dataset in relation to the methods of cleaning. As a first step, cleaning methods were classified into wet, dry and mixed with the aim of identifying cleaning preferences in relation to object classes and material types. Nominal data such as that collected need further manipulation before analysis can be carried out. The methods used varied from simple descriptive statistics like frequency tables to factor analysis, principal component analysis (PCA), multiple correspondence and cluster analysis which revealed cleaning patterns in the collection.

The preliminary results showed there are three main outliers in the general pattern: paintings, a group of metals and textiles. It is interesting to note that in the case of metals the pattern is observed only on objects belonging to social history collections. Each of the three groups is treated

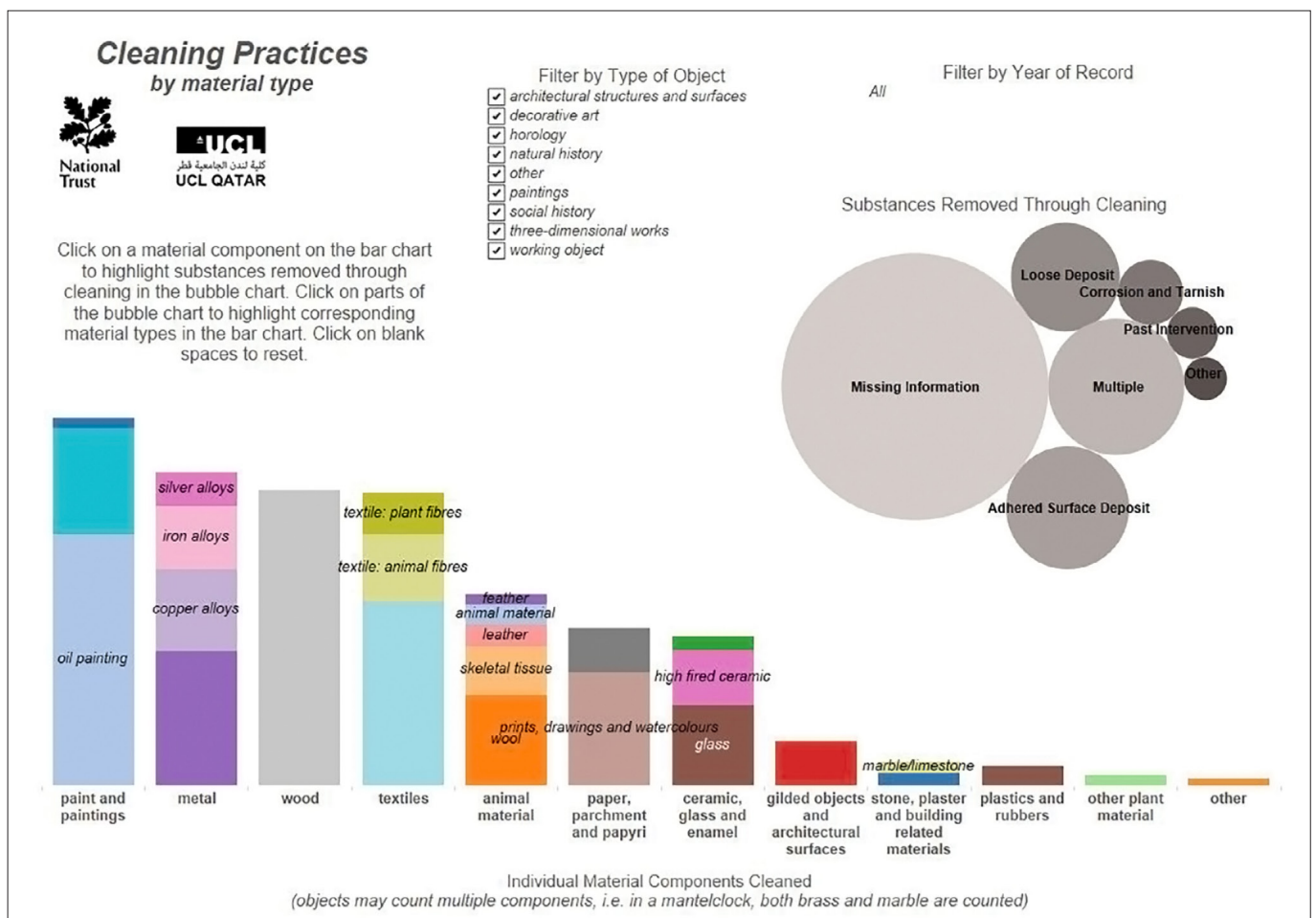


Figure 3. The bar chart demonstrates which materials have been given most attention, while clusters indicate the type of layer or soiling removed from objects, insofar as conservators' descriptions distinguish between them

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

preferentially with a specific type of treatment, with paintings cleaned mostly with wet methods (which include solvent cleaning), textiles with dry methods and metals in social history objects with mixed methods. Figure 5 shows the results of the PCA analysis for the cleaning methods in relation to the material. Additional analysis was carried out to see the relationship between the type of substance removed (e.g. loose deposits, adhered deposits, past interventions, corrosion, multiple, other) and the material cleaned. The patterns emerged were again related to cleaning of paintings and textiles.

**DISCUSSION: DEALING WITH UNCERTAINTY IN
CONSERVATION DATA**

Museums employ a variety of databases, some developed in-house as a simple archive that provides information for specific objects without foreseeing their use for practices followed within the entire collection. Databases populated by different people can be inconsistent in the way information is recorded. By demonstrating the potential of the information stored in these records it is hoped that studies such as the one presented here will encourage museums to use a more structured format for data input to facilitate analysis. Agreement on a single data standard seems unrealistic even amongst the conservators working for the same institution, particularly where the wider scope of these data is not yet envisaged.

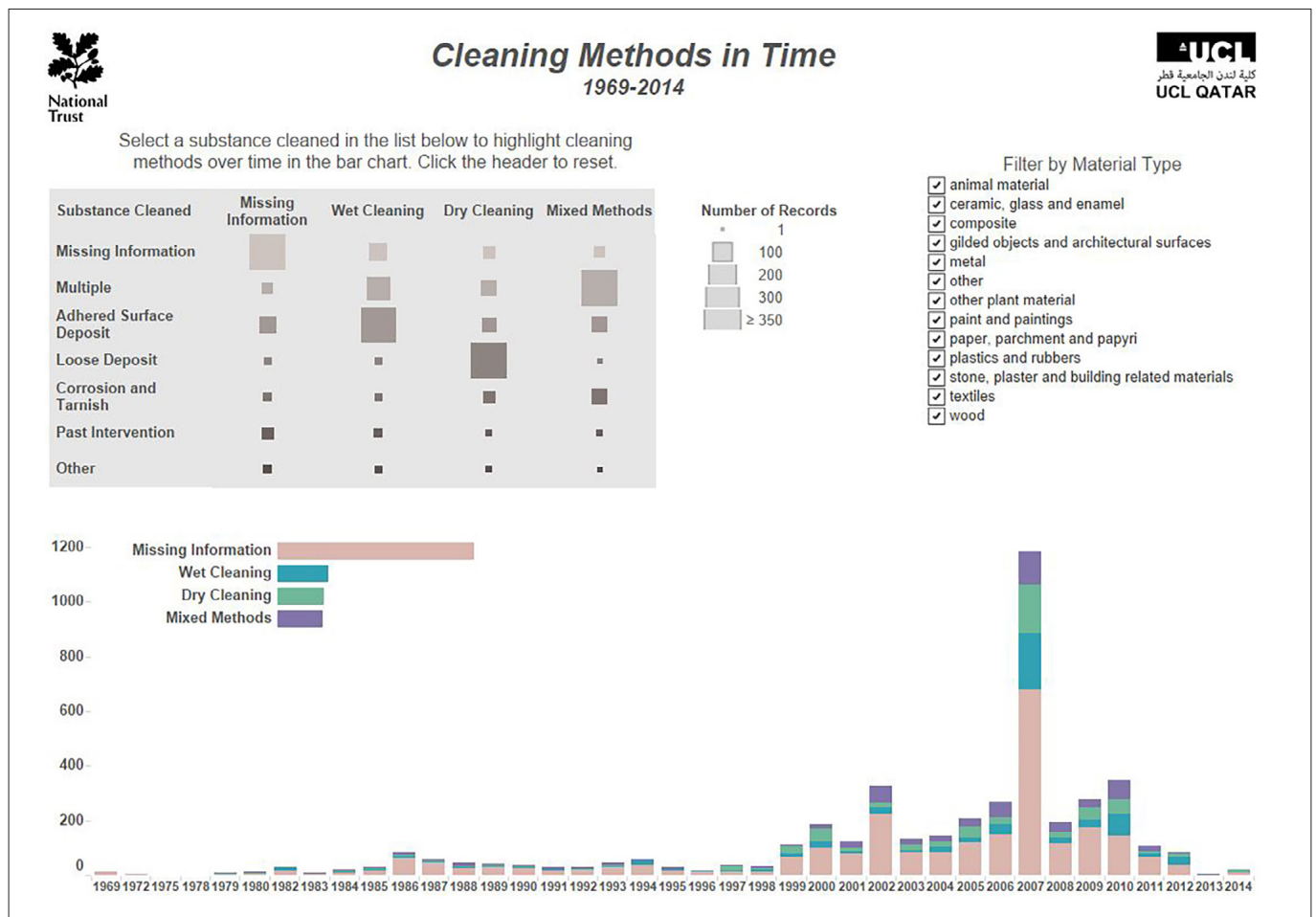


Figure 4. The chart shows predominance of dry cleaning until 2000, when greater resources were invested in conservation allowing for more complex treatments to be undertaken

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

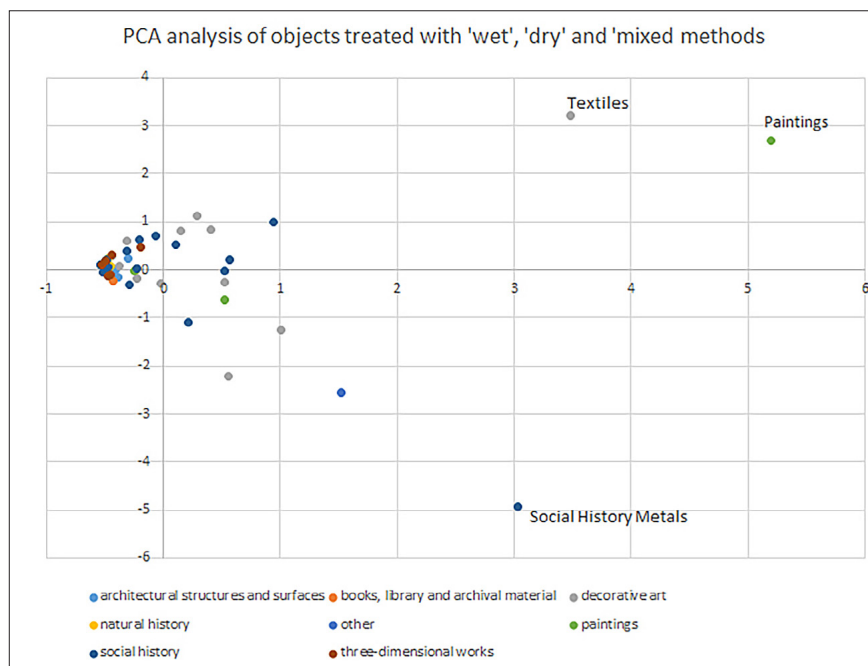


Figure 5. The graph shows the results of the PCA analysis for the cleaning methods in relation to the material

In this study, data collection and clean-up found that records had missing information (Figure 4). Data is not always compliant with institutional standards for documentation: for example, despite the condition code system, which was implemented through CMS in 2005, condition rating systems used by other institutions or individuals with which freelance conservators were evidently more familiar are sometimes used alongside the official code system. The majority of the reports in CMS were PDF, Word and Excel file attachments which needed to be opened one by one, requiring a considerable time investment.

Some conclusions of this study may be drawn only after comparison with datasets from different organisations and countries (for example, the British Museum and the National Trust records for all properties). Nevertheless, data mining techniques applied to National Trust data allowed trends and relationships to be readily identified, such as the prioritisation of certain types of materials, and rates of activity over time. Clustering suggests causal relationships, from the more obvious such as the link between substrate and the type of cleaning used in different types of objects, to the less obvious such as specific groups of objects (e.g. Metals in Social History Collections) which appear as outliers in relation to the methods chosen. The consistency of the data with known events such as particular projects, organisational changes and financial investment reinforces the robustness of the coding methodology and analysis. The quality of documentation has improved over time, from short summaries in the 1980s to more detailed descriptions of the nature of removed materials in more recent years, reflecting higher standards and specification of documentation as the profession develops. Further insight, such as the rationale for cleaning and other conservation decisions of both conservator and client, or the relationship between agent of deterioration and damage, could be explored by further coding of the data. This would be facilitated by the use of standardised terminology and fields with drop-down predefined options. It should also be noted that the greater

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

the degree of prescription, the more onerous the work of documentation, and the less time there is for insight and judgement. Nevertheless, identifying questions that would help the development of professional practice as well as conservation management could influence future database design. Thus far, the study has shown that data mining techniques have the potential to identify more relationships than the original reports overtly describe.

Several software programmes are available to assist conservation professionals in the analysis of data, and no doubt more will appear in the near future. This study has served to compare and contrast different analytical methods. The principal issue was encountered in the preliminary phase of data retrieval, organisation and cleaning, which proved time consuming for a relatively small set of data. Automated data mining techniques might help expedite these steps, thus extending the possibility of analysis to larger and more complex data sets. An example of this is supervised machine learning, which uses classification based on examples. Unsupervised learning is done by clustering, when the class labels are unknown. In supervised learning the computer needs a significant number of examples before it learns the relevant terminology. Accuracy is dependent on the question and the data collected; however, there are methods that take into account the missing information which cannot be otherwise possible.

The analysis of the data collected is still in progress. The initial results looked at the overarching categories, while further analysis will look into the detailed materials and methods used. Additional investigation is needed to determine whether cleaning happened as a consequence of another treatment or whether it was the primary aim of treatment, which is one of the main aims of the Coming Clean project. Cleaning is a treatment step which might not be recorded in detail if it is not the primary aim of the treatment carried out. Yet it happens and it is a laborious activity. Qualitative analysis of treatment records and discussions with key conservation advisors from the National Trust will shed light into these questions.

The use of these data can be a great resource for collections care managers so they can understand where the time and money are spent and patterns of professional behaviour within an organisation. The information extracted can also be used to inform evidence-based research strategies in relation to conservation and heritage science. For example, identifying the most prevalent types of treatments and related uncertainties can assist organisations in deciding where their research focus should be. This will lead to research of greater relevance and impact in the field as it will be based on evidence-based needs assessment (see also Heritage and Golfomitsou 2015). Another interesting application is to look into treatment cycles and whether certain treatments lead to others and why. The latter can be done by looking at the correlation in the occurrence of different treatment types and the underlined decisions taken.

CONCLUSION

Conservation records are considered crude data for which a large sample size is required in order to reach meaningful conclusions. While a wide sample of data allows more accurate conclusions to be drawn, this has

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

to be accompanied by closer examination of smaller samples containing more detailed information. Analytical methods used in other information management contexts have allowed the *what, when, how* and *who* of conservation practice relating to cleaning at the National Trust to be investigated. Statistical methods painted an overall picture of the conservation work carried out over time, and extracted detail on the methods employed. This was essential before moving on to the next phase in which the authors will address *why* certain methods were employed. Focusing on the sections of conservation records that describe justification for treatment, the rationale of conservation choices will be explored using text mining and natural language processing techniques. Although variable in nature, conservation records can still be analysed using data mining techniques to provide a wealth of information regarding treatments. The approach presented here was found to be useful in reviewing decision-making processes in heritage institutions and informing future conservation evidence-based research strategies.

ACKNOWLEDGEMENTS

The authors would like to thank Catherine Dillon, Stefan Michalski, Adrian Heritage, Dean Sully, Cymbeline Storey, Maria Carmen Vida, Mike Charlton and Alison Heritage. Finally, we would like to thank UCL Qatar for funding this project and Professor Thilo Rehren for his continued support.

REFERENCES

- GOLFOMITSOU, S. and J.F. MERKEL. 2004. Synergistic effects of corrosion inhibitors for copper and copper alloy archaeological artefacts. In *Metal 04: Proceedings of the International Conference on Metals Conservation, National Museum of Canberra ACT, 4–8 October 2004*. Canberra: National Museum of Australia.
- HERITAGE, A., C. ANUZET, E. ANDERSSON, and C. AN TOMARCHI. 2014. The ICCROM Forum on Conservation Science 2013: A collaborative partnership for strategic thinking. In *ICOM-CC 17th Triennial Conference Preprints, Melbourne, 15–19 September 2014*, ed. J. Bridgland, art. 1903. Paris: International Council of Museums.
- HERITAGE, A. and S. GOLFOMITSOU. 2015. Conservation science: Reflections and future perspectives. *Studies in Conservation* 60(S2): 2–6.
- HODDER, I. and C. ORTON. 1976. *Spatial analysis in archaeology*. Cambridge: Cambridge University Press.
- KARSTEN, I., S. MICHALSKI, M. CASE, and J. WARD. 2012. Balancing the preservation needs of historic house museums and their collections through risk management. In *The Artifact, Its Context and Their Narrative: Multidisciplinary Conservation in Historic House Museums. Proceedings of the ICOM-DEM HIST and ICOM-CC Working Groups Sculpture, Polychromy, & Architectural Decoration; Wood, Furniture, & Lacquer; and Textiles, Getty Research Institute, Los Angeles, 6–9 November 2012*, eds. K. Seymour and M. Sawicki, 1–12. Canadian Conservation Institute (http://www.icom-cc.org/ul/cms/fck-uploaded/documents/DEM HIST%20_%20ICOM-CC%20Joint%20Interim%20Meeting%202012/10-Karsten-DEM HIST_ICOMCC-LA_2012.pdf).
- LITHGOW, K., S. STANFORTH, and P. ETHERIDGE. 2008. Prioritizing access in the conservation of National Trust collections. *Studies in Conservation* 53(S1): 178–85.
- LITHGOW, K. AND H. LLOYD. 2017. Direct preventive conservation – Using information from the past to prevent small issues in the present from becoming bigger problems in the future. In *ICOM-CC 18th Triennial Conference Preprints, Copenhagen, 4–8 September 2017*, ed. J. Bridgland, art. 1513. Paris: International Council of Museums.
- SAWDY, A. and C. PRICE. 2005a. Salt damage at Cleve Abbey, England. Part I: A comparison of theoretical predictions and practical observations. *Journal of Cultural Heritage* 6: 125–35.

DOCUMENTATION

**OFF THE RECORD: USING DATA MINING
TO REVIEW DECISION MAKING IN
CONSERVATION PRACTICE**

- SAWDY, A. and C. PRICE. 2005b. Salt damage at Cleeve Abbey, England. Part II: Seasonal variability of salt distribution and implications for sampling strategies. *Journal of Cultural Heritage* 6: 269–75.
- SHENNAN, S. 1997. *Quantifying archaeology*. Edinburgh: Edinburgh University Press.
- SUENSON-TAYLOR, K., D. SULLY, and C. ORTON. 1999. Data in conservation: The missing link in the process. *Studies in Conservation* 44(3): 184–94.
- REEDY, T.J. AND C.L. REEDY. 1988. *Statistical analysis in art conservation research*. Marina del Rey, California: Getty Conservation Institute.
- ROY, A., S. FOISTER, and A. RUDENSTINE. 2007. Conservation documentation in digital form: A continuing dialogue about the issues. *Studies in Conservation* 52(4): 315–17.
- XAVIER-ROWE, A., C. FRY, and B. STANLEY. 2008. Power to prioritize: Applying risk and condition information to the management of dispersed collections. *Studies in Conservation* 53(S1): 186–91.
- WALLER, R.R. 2003. *Cultural property risk analysis model: Development and application to preventive conservation at the Canadian Museum of Nature*. Goteborg, Sweden: Acta Universitatis Gothoburgensis.

How to cite this article:

Golfomitsou, S., F. Ravaioli, C. Tully, G. McArthur, and K. Lithgow. 2017. Off the record: Using data mining to review decision making in conservation practice. In *ICOM-CC 18th Triennial Conference Preprints, Copenhagen, 4–8 September 2017*, ed. J. Bridgland, art. 0202. Paris: International Council of Museums.