The Role of Heritage Data Science in Digital Heritage

Alejandra Albuerne[1[0000-0001-6444-9120]✉], Josep Grau-Bove[2] and Matija Strlic[3]

[1] Institute for Sustainable Heritage, University College London, London, United Kingdom
a.albuerne@ucl.ac.uk
[2] Institute for Sustainable Heritage, University College London, London, United Kingdom
[3] Institute for Sustainable Heritage, University College London, London, United Kingdom

Abstract. The advance of all forms of digital and virtual heritage alongside numerous heritage science and management applications have led to the generation of growing amounts of *heritage data*. This data is increasingly rich, diverse and powerful. To get the most out of heritage data, there is an evident need to effectively understand, manage and exploit it in a way that is sensitive towards its context, responding to its singularities, and that can allow heritage to keep up with global changes regarding expansion of digital technologies and the increasing role of data in decision making and policy development. Through conversations with industry and academia, as well as through their personal research in the field of cultural heritage, the authors have identified a need for enhanced training for data scientists to prepare them for working in the heritage sector. This paper first proposes a definition of the term *heritage data*, so far missing from the literature, and then presents the academic rationale behind the identified need for targeted training in data science for cultural heritage.

Keywords: Heritage Data, Data Science, Cultural Heritage.

1 Introduction

The relevance of digital technologies and extent of use of data in today's world is growing at a fast pace. The field of heritage is no exception, having witnessed a vast expansion of digital heritage over the past two decades. The increasing amounts of data being generated give rise to many new possibilities and opportunities, as well as many challenges. There is an evident need to effectively understand, manage and exploit data in the cultural heritage context, not only in order to take advantage of possibilities and manage challenges, but also in order to ensure the heritage sector can keep up with the increasing role of data in decision-making and policy development [1]. This requires data science specific skills which are not part of the traditional skills set of the cultural heritage field.

To what extent are these skills available in the sector is explored in this paper through interrogation of the sector by means of questionnaires and semi-structured interviews, as well as the authors' own professional experience.

The Institute for Sustainable Heritage of University College London (UCL ISH), in collaboration with partners, conducted in 2017 a survey of training provision in heritage

science, where 262 active professionals (heritage scientists, curators, conservators and other industry professionals) were requested to provide views on current training offer. This survey was carried out as part of the initiative to set up the European Research Infrastructure for Heritage Science (E-RIHS). Approximately 10% of the questionnaire focused on data science, heritage data analysis and digital training provision.

Between March and May 2018, UCL ISH conducted a series of semi-structured interviews with experts from leading heritage organizations that included The British Library, English Heritage and Historic Environment Scotland in the UK and The Library of Congress in the US, as well as data science experts from the Alan Turing Institute (UK). The interviews focused on the need for data science skills in the field of cultural heritage, addressing aspects such as existing technical challenges, available training and difficulties in recruitment of data scientist to work in heritage institutions.

This evidence is complemented with a desktop review of available university programmes in the UK and in leading European and North-American universities.

A review of this evidence strongly suggests there is a need for targeted training for data scientists to equip them to meet the needs of the heritage sector.

2 Heritage Data

Data science is the set of knowledge and processes that facilitate the creation of data products [2]. It operates with and transforms digital data. The application of data science in the remit of cultural heritage therefore calls for a definition of heritage data. The term heritage data in the remit of cultural heritage first appears in the technical literature in 2006 [3]. It is used to broaden the scope of the existing term archaeological data that had been used in relation to digital applications since the late 20th century, e.g. [4], [5].

Although there is currently no formal definition, heritage data is often used in reference to:

- data generated from the documentation of heritage, be it capture/acquisition, processing/analysis or visualization, e.g. [6], [7];
- ontology of heritage for applications in data management, archiving and web-based dissemination [8], [9].

Maricevic [1] suggests there is at present a notion of heritage data community that is experienced by those involved in the subject, but highlights the gap that exists within the sector in understanding whether there is a common theme of heritage data that can ground this perceived shared sense.

Embracing Maricevic's observation and supported by views expressed by different stakeholders in the heritage sector (see sections 1 and 3), we propose defining heritage data as a comprehensive term that leaves out no exceptions and includes both data *as* heritage and data *about* heritage.

2.1 Data *as* heritage

In their *Charter on the preservation of digital heritage* [10], UNESCO recognises digital contents as heritage that may have both lasting value and significance and must therefore be preserved under the same premises as other forms of cultural heritage for future generations.

The charter refers to a wide range of digital materials that can be identified as heritage, from visual materials, such as moving and still images or graphics, to audio or text. It also makes specific mention to less evident forms of digital objects that may also be considered heritage, including software, web pages or databases. The range of digital materials subject to being identified as heritage is indeed growing.

These digital materials are comprised of digital data and this digital data must therefore also be recognised as heritage.

2.2 Data *about* heritage

Data related to the documentation, conservation, management and interpretation of heritage constitutes an extensive and increasingly significant body that requires growing amounts of capabilities and resources for its management and exploitation.

This comprehensive definition of heritage data comprises a broad range of data, both quantitative and qualitative, of different origins and for different uses. The commonality lies in the fact that this data exists to create value for the field of cultural heritage. As such, its handling must reflect the values, significance, integrity, ethics, authenticity and other particularities of the heritage sector it belongs to. Alongside this, there is the singular need for longevity that is characteristic of the sector.

Longevity, in terms of digital data, requires reliability and planning at all stages of the data pipeline/cycle: acquisition, analysis - visualization - storage - repurposing - curation and conservation; be it in the form of data standards, archiving and sharing practices, etc. These principles can be relevant for and common to all types of heritage data.

Furthermore, there is a common thread of multidisciplinarity throughout the heritage sector that introduces its own specific challenges, such as bringing together expertise from diverse fields of arts, humanities, social sciences, science and technology.

These commonalities are strong justifications for a holistic definition of *heritage data*. The idea that data science can extract and create new value from data further supports this proposition for a unified theme of *heritage data* and the development of a common set of interrelated skills to support it.

3 A need for targeted training

3.1 Sector consultations

Consultations conducted by UCL ISH between 2017 and 2018 among the heritage community have yielded findings strongly suggesting that, alongside the need for data science skills, in the cultural heritage sector there is need to enhance the training of data

scientists to prepare them for those specific characteristics and challenges of the sector that set it apart from more conventional applications of data science.

As part of the Preparatory Phase project supporting the establishment of the European Research Infrastructure for Heritage Science (www.e-rihs.eu), a stakeholder questionnaire was conducted with 282 respondents from heritage and research institutions. Of the academic courses recently undertaken, only 6% focused on heritage science data analysis, visualization, use and reuse; 3% on digitalization/digital heritage; and 1% on data science. One of the conclusions of the questionnaire was to increase the provision of courses specifically focusing on digital and data skills.

Semi-structured interviews were conducted between March and May 2018 with experts from selected heritage organizations with interest and capabilities in data science (see section 1). The key observations extracted from these interviews are:

- The needs for data science skills are growing in the consulted organizations, as they embrace digital technologies and their collections of heritage data grow. Including data scientists in the work force of heritage organizations is becoming an increasingly regular occurrence.
- Finding data scientists to work successfully in the cultural heritage sector fully embracing its goals, values and ethics is perceived as a challenge by the heritage organizations consulted.
- The need for longevity of most heritage data poses significant challenges as digital technologies advance at a growing pace and the sector attempts to keep up with the changes. This results in specific needs of data migration and recovery that are unique in the remit of data science.
- The lack of heritage data standards is problematic when planning data acquisition, management and curation.
- The experts consulted were unaware of any training currently available that comprehensively addresses the range of data science needs experienced in the sector.

Further to these findings, the recent Arts and Humanities Research Council (UK) report *Heritage And Data: Challenges And Opportunities For The Heritage Sector* [1] identified significant challenges that need to be addressed in heritage data governance.

Outputs on existing training from the E-RHIS survey and from the interviews of sector experts have been complemented with a review of available university programmes in the UK and in leading European and North-American universities. The conclusion is that the diverse skills needed to address the data science challenges of the cultural heritage sector are currently not being taught comprehensively.

The context is important for the work of the data scientists applying their technical skills to cultural heritage. Such data scientists must understand heritage values and how heritage enriches society. The sector is looking for data scientists that have capabilities that go beyond classical technical skills and into skills that are conventionally linked to social sciences or humanities. In particular, these include a critical approach, a consideration of societal issues and strong communication skills that enable them to communicate with multidisciplinary teams and diverse audiences.

Conventional data science masters programmes currently on offer fall short in developing these skills, which are excluded from the programmes' curricula. This has

consequences for the heritage sector, which expresses difficulty in finding professionals with these broader sets of skills.

3.2 Addressing the needs

Heritage data diversity.
Heritage data is diverse. It encompasses both qualitative and quantitative data generated through many different forms of technology and for very different purposes. Furthermore, there is digitally born data, e.g. through social media and other digital social interaction, which is increasingly collected by heritage institutions and that can offer, among other benefits, exciting opportunities for engaging with the community. This data, together with digitally-born heritage, represents the core of digital collections that require specific curation and conservation approaches. Additionally, there is the need to address the variety of data, both qualitative and quantitative, that is generated around heritage through analysis and measurement, imaging and surveying, or through citizen science and publicly sourced data. This requires a variety of highly technical data science skills.

Furthermore, it is crucial that data scientists working in cultural heritage should have the capacity to be critical about the provenance, quality, accuracy and bias of the data they are managing. With data being collected in growing quantities, the successful interpretation of these large heterogeneous data sets requires expert judgement by professionals with deep knowledge of the heritage domain, beyond data science specific knowledge. This view has been expressed in [1], echoed in interviews of sector experts and is shared by the authors.

Understanding the heritage data pipeline.
It is clear that there is a well-established and growing need for heritage researchers and scientists with in-depth understanding of the heritage data pipeline or lifecycle: acquisition - analysis - visualization - storage - repurposing – access - curation and conservation. While there are individual courses available in the UK and the US that address specific aspects of this pipeline, there is currently no Masters programme on offer that addresses the complexities of data acquisition, exploitation, management and conservation in cultural heritage.

There are graduate programmes available that address one or more of the aspects of the heritage data pipeline. Examples include programmes on information science, which generally cover some acquisition and analysis processes, visualization, access, repurposing, etc., the limitation being that the focus is solely in information, leaving many forms of heritage data out. Archaeological programmes frequently address other types of heritage data, such as GIS or 3D scans and models. Heritage science programmes address the acquisition and analysis of heritage conservation data. Digital heritage programmes address multiple aspects of data acquisition and management utilizing state of the art technologies.

All these types of programmes, however, focus only on a fragment of the pipeline or lifecycle of heritage data. Furthermore, they do not attempt to train data scientists: they

train other specialisms, be these archaeologists or heritage scientists, who may gain an insight into data science but frequently lack the deep core knowledge and skills that could be applied to diversified problems.

Big Data.
The ability to operate with Big Data is also of great importance, as the increasing amounts of data being generated, directly and indirectly as by-products, call for new and innovative strategies for analyzing, managing and preserving data.

Integrating data science in the heritage interdisciplinary work.
In all the above there is abundant justification for integrating data scientists in the core of the cultural heritage discipline. This research has documented that currently heritage organizations struggle to successfully integrate data science into the core of their work, as they encounter a challenge in meaningfully engaging data scientists in the complexities and multidisciplinarity of heritage challenges, thus facing a disconnect between the objectives of the sector and the contribution of data science. As the role of digital heritage grows in the sector, the need for these skills will also grow.

Targeted training is needed to equip data scientists for becoming integral parts of the heritage sector. Such training should explore the full heritage data pipeline or lifecycle and should equip professionals with specific data science skills such as data visualization, machine learning, data migration, image processing, etc., through direct applications to heritage data and digital heritage in order to provide an understanding of the technical challenges around heritage. As well as technical contents, a deep insight of the context must be provided to include underpinning concepts and challenges such as heritage value and significance, integrity and authenticity, ethics, etc.

This comprehensive training can best prepare data scientists to be fully embedded in the work of heritage organizations from the early days. Without this training, however, the sector is at risk of falling behind in some of the major developments that the scientific community is enabling worldwide. The research carried out to date has further documented that this training is currently unavailable. An effort is required to develop courses and academic programmes that can address this need. Such effort will entail continued collaboration between academia and heritage organizations and practitioners, as well as networks such as E-RIHS, to develop targeted learning objectives and course contents and to think creatively about possible future challenges and opportunities pertinent to heritage data science. UCL ISH is pursuing further research on this subject, continuing their consultations with the heritage community to develop detailed learning objectives for heritage data science with the ultimate objective of strengthening the interdisciplinary heritage sector.

4 Conclusions

Heritage data is defined as a comprehensive term that encompasses both data as heritage and data about heritage, drawing on the commonalities of this data: it is underpinned by

heritage values, significance, integrity, ethics and authenticity; it is characterised by a need for longevity; and it is the product or the object of multidisciplinary work.

The needs of the cultural heritage sector for data science training and skills have been explored by means of questionnaires and semi-structured interviews, as well as the authors' own professional experience. These consultations have been complemented with a desktop review of available university programmes in the UK and in leading European and North-American universities.

Findings strongly suggest that there is a need for data science skills that are currently not easy to find in the market and that are not being taught comprehensively. The sector needs data science professional who are familiar with both the specific technical requirements of cultural heritage data and the particularities of the cultural heritage sector.

With the current expansion of heritage data, there is a growing need for professionals with in-depth understanding of the heritage data pipeline or lifecycle in order to exploit, manage and preserve this data and who can exert expert judgment to assess the provenance, quality, accuracy and bias of heritage data. Simultaneously, these experts must understand the heritage sector, its values, principles and ethics and have skills that go beyond classical technical skills and into skills that are conventionally linked to social sciences or humanities, such as critical analysis and communication.

Providing the cultural heritage sector with skilled data scientists that fully embrace its singularities will be crucial for enabling the sector to keep up with global changes regarding expansion of digital technologies and the increasing role of data in decision making and policy development. Continued work is being undertaken at UCL ISH as a collaboration between academia and the professional heritage sector to develop learning objectives and course contents aimed at creating new training opportunities that meet the data science needs identified in the heritage sector.

5 Acknowledgements

We are grateful to all interviewees and questionnaire respondents for their generous collaboration and to Yujia Luo for her last-minute help with Mendeley.

References

1. R. Harrison, H. Morel, M. Maricevic, and S. Penrose, "Heritage and Data: Challenges and Op-portunities for the Heritage Sector."
2. M. Loukides, "What is Data Science?," *O'Reilly Media*, 2011. .
3. E. Meyer, P. Grussenmeyer, J. P. Perrin, A. Durand, and P. Drap, "Integration of heterogeneous cultural heritage data in a web-based information system: a case study from Vianden Castle, Luxembourg," in *CAA 2006 Proceedings.*
4. P. Reilly and S. Rahtz, "Archaeology in the Information Age: A Global Perspective. One World Archaeology 21." London: Routledge, 1992.
5. J. D. Richards, "Recent trends in computer applications in archaeology," *J. Archaeol. Res.*, vol. 6, no. 4, pp. 331–382, 1998.

6. M. L. Vincent, V. M. L.-M. Bendicho, M. Ioannides, and T. E. Levy, *Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data*. Springer, 2017.
7. N. Lercari, E. Shiferaw, M. Forte, and R. Kopper, "Immersive Visualization and Curation of Archaeological Heritage Data: Çatalhöyük and the Dig@ IT App," *J. Archaeol. Method Theory*, pp. 1–25, 2017.
8. "http:/www.heritagedata.org/blog/vocabularies-provided/ Date accessed." .
9. L. Klic, J. K. Nelson, M. C. Pattuelli, and A. Provo, "Florentine Renaissance Drawings: A Linked Catalog for the Semantic Web," *Art Doc. J. Art Libr. Soc. North Am.*, vol. 37, no. 1, pp. 33–43, 2018.
10. UNESCO, "Charter on the preservation of digital heritage," 2003.