CHAPTER 32
Heritage data science

Scott Allan Orr

ABSTRACT

The so-called ‘Data Revolution’ is rapidly transforming society, including the heritage sector. Building on the more-established area of heritage science, a framework for heritage data science is proposed as a transdisciplinary field that employs data-driven approaches with critical reasoning within the heritage domain, in awareness of its unique and pressing challenges, to inform engagement with heritage and its interpretation and long-term management. Several open challenges within the field are discussed, including data quality and integrity, transdisciplinary, and education.

Introduction

‘The Data Revolution’ (Kitchen, 2014) is rapidly changing nearly every aspect of society. The use of data and data-driven approaches underpins infrastructure, healthcare, education, and economic decision making. These approaches, regardless of specific application, hold the promise of identifying and processing patterns and trends beyond the capabilities of traditional approaches humans use to processing information to inform decision making.

The heritage sector is not exempt from these developments. Decision making increasingly relies on scientific activity to provide evidence, which is typically underpinned by robust experimental procedures and data. A notable area of earlier development is digital humanities (Schreibman, et al., 2008), a field at the intersection of digital technologies and the humanities tradition (Terras, 2011). ‘The Data Revolution’ is further exacerbated within the heritage sector

by the vast heritage assets that are ‘born digital’, with no physical analogue or manifestation (Palfrey and Gasser, 2011), which have accelerated the need to address the sustainability and management challenges posed by these resources.

While the generation and use of data is widespread and well-established, the concept of ‘data science’ provides a novel framework in which to understand its creation and use. In casual use, the terms statistics and data science may be used interchangeably. The astute observer might identify a distinction of scale or scope, specifying that something becomes data science when the data is ‘big’. Or, they might believe that machine learning must be involved to be considered data science. These distinctions are artificial: statistical science and data science can be based on nearly any scale of data and may or may not include machine learning as part of their toolkit. These misperceptions of data science are rooted in the origins of data science within mathematical communities and primarily emphasising data analysis (Tukey, 1962). Both observers in the previous example have omitted two crucial things: context and critical reasoning.

Theory

Data science is a transdisciplinary field that incorporates several relevant bodies of knowledge and disciplinary approaches, including but not limited to statistics, informatics and communications technology (ICT), management, and sociology (Cao, 2017). These approaches are used in combination on the basis of three interrelated components: data, domain (context), and thinking (innovation). The heritage domain, with its unique and pressing challenges and great diversity of value frameworks and disciplinary epistemologies, gives rise to the emerging field of *heritage data science* (Albuerne et al., 2018).

The data pipeline

The data pipeline is a set of activities that enable data-driven decision making:

- Conceptualisation, including design, planning, and stakeholder engagement
- Acquisition, including methodological design, data collection, and documentation/recording
- Processing, including data cleaning, manipulation, synthesis, and conversion
- Analysis, including statistical summaries, algorithmic processes (e.g., machine learning)
- Visualisation and interpretation, typically including graphical representation of analysis, written descriptions, and discussion
- Curation and long-term management, including data documentation and storage

Although the components of the pipeline are conceptualised as a sequential procedure, they are strongly interrelated. For example, the nature of acquisition (how, when, where, and by whom) have significant implications for interpreting the analysis. If the analysis is ignorant of the methods, context, or design of data acquisition there is a risk that a limitation or bias may not be identified.

The components of the data pipeline are most effectively implemented when iterated. The action of processing, analysing, visualising and interpreting data rarely addresses all relevant aspects of the scenario. At each stage of the data pipeline or a data-driven project, it should be reassessed whether further work 'up the pipeline' is required, forming several interwoven feedback loops.

Heritage as a domain

Understanding the domain of a data science project is essential to formulating and undertaking data-driven decision making. The domain provides context to any data science approach: it determines the subject of investigation and identifies the relevant challenges and questions to be addressed with data science. Within heritage data science, the data pipeline sits within the

heritage domain: it is informed by heritage challenge(s) and aims to produce output that is relevant to the priorities of heritage stakeholders.

The 'subject' of heritage data science may be an individual heritage typology, or a set of heritage typologies. Heritage data science is not limited to the either the immaterial or material, as it transcends the false dichotomies of cultural/natural and tangible/intangible heritage (Fredheim & Khalaf, 2016). Thus, heritage data science is undertaken on material and/or immaterial culture to which a society ascribes value (Vecco, 2010). Data science methods are particularly adept at handling large sets of information, especially those that contain conflicting information and uncertainty and/or are based on subjectivity, perception, or belief. Thus, heritage data science is particularly useful for addressing complex challenges within the heritage domain.

The diverse interests of stakeholders in the heritage domain necessitate a wide range of activities within heritage data science. Building on an established framework in heritage science (NHSF, 2018), these activities can be broadly classified into:

- Interpretation: furthering understanding of heritage
- Engagement: enabling and enhancing access to heritage
- Management: informing the stewardship of heritage, including but not limited to storage and maintenance

One of the grand challenges for heritage data science is to reconcile that the conceptualisation and implementation of a project are often rooted in several disciplinary epistemologies. As a transdisciplinary field, heritage data science incorporates methods and frameworks from several different traditions of study. This is exacerbated by the diversity of participants and applications of heritage data science. Some of these challenges are practical and can be addressed with the implementation of practical steps. For example, a widespread challenge to

transdisciplinary working is a lack of shared terminology, or understanding of terminologies used by different disciplines (Tress at al., 2007). These can be addressed by exercises, such that as proposed by the computer scientist and noted internet pioneer Jim Gray, in which the task to formulate *20 Questions* relevant to a societal challenge assists in the process of normalising understanding of terminology and priorities (Gray, 2004). Other challenges, such as contrasts between disciplines in how claims are proved or substantiated, remain long-standing open challenges.

Thinking

Data science 'thinking' (Cao, 2018) embodies the distinction between data science different from existing developments in statistics and information science. Data science thinking emphasises the role of human intelligence by involving human intuition, belief, expectations, evaluation, and expertise into decision making processes. In addition to the ability to collect, process, and analyse vast quantities of data, which can in some cases be routine tasks or trivial processes, heritage data scientists must also be imaginative and employ qualitative and critical reasoning.

Heritage data science thinking must critically assess the objective of study against the priorities of stakeholders. Heritage data science should actively address an open challenge within the engagement with, interpretation, or management of heritage. The nature of how this is undertaken depends on the context within the heritage domain and the relevant stakeholders, but should broadly seek to positively enhance the value of heritage and the benefits derived from it. These can be represented by one or several objectives, including (in no specific order):

Optimise Economise Reason

Automate	Imagine	Explain
Communicate	Calculate	Evaluate
Innovate	Critique	Prove
Demonstrate	Reduce	Predict

The appropriate objective(s) being determined from stakeholder priorities.

A framework

Heritage data science sits at the intersection of the heritage domain, critical and imaginative thinking, and data (Figure 32.1).

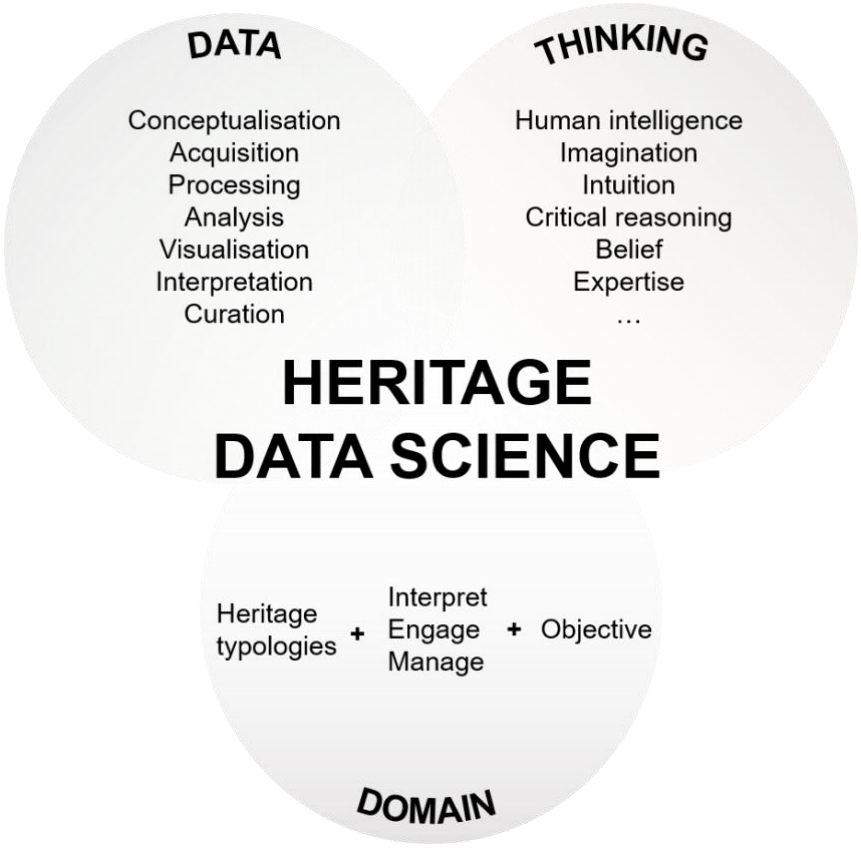

Figure 32.1. A framework for heritage data science, combining data, thinking, and the heritage domain.

Crucially, the data pipeline is integrated within this framework. This is fundamentally different than the model in which a challenge is generated within the heritage domain which inputs into the project, at which point the data pipeline is implemented by those with the relevant technical expertise. Although this may produce implications for the interpretation, engagement, and management of heritage, it is limited by its linearity. As a transdisciplinary field, the heritage domain (and relevant stakeholders) should be involved throughout the duration of a heritage data science project as a realisation of participatory science.

Heritage data

Heritage data is an inclusive term that includes data *as* heritage and data *about* heritage (Albuerne et al., 2018). The unified term is useful to identify common challenges and opportunities for heritage data: it is rooted in established concepts within sustainable heritage of value, significance, integrity, ethics and authenticity; it imposes requirements for longevity; and it is the subject and/or output of transdisciplinary work.

Data as heritage

The UNESCO *Charter on the preservation of digital heritage* (2003) recognises digital assets as heritage that have lasting value and significance and should therefore be safeguarded under the same premises as other forms heritage. The charter identifies a diverse range of types of digital materials that can be heritage, including texts, databases, still and moving images, audio, graphics, software and web pages, among a wide and growing range of formats.

Data *about* heritage

There is an ever-increasing amount of data related to the provenance, conservation, management and interpretation of heritage. This data can be both qualitative and quantitative,

scientific or humanistic, and come from diverse sources, and serve several different purposes. The unifying concept for these data is that they exist to produce, promote, and sustain value for heritage; thus, data about heritage is embedded in concepts of authenticity, ethics, integrity, and values, just as heritage to which they are relevant are embedded.

Sustainability of heritage data

Heritage data requires the same diligence as other forms of heritage with regard to sustainable management. This is especially true as heritage data are often ephemeral, requiring purposeful maintenance and management to be retained. Accessing heritage data in proprietary formats, specifically the challenges associated with the maintenance of software required to access them, remains a significant challenge in heritage data science. This challenge is exacerbated when the proprietary format has in itself embedded heritage value, such as legacy video game platforms (see Eklund et al. 2019).

A heritage data science paradigm

Data science is argued to be an *exploratory* mode of science (Hey, 2009): a 'fourth paradigm' of undertaking scientific enquiry is proposed that is fundamentally different from its three predecessors that are, respectively, rooted in empirical evidence (observation), the scientific method (hypothesis testing), and computational science (scaling up analysis based on computation). In the fourth paradigm, data is big, abundant, and rich. In contrast to previous approaches to science, data is not collected based on a hypothesis; in the fourth paradigm data is collected first, from which patterns, insights, and information are extracted. Collect first, ask questions second.

The 2006 House of Lords Science and Technology inquiry into science and heritage (House of Lords Science and Technology Committee, 2006) was grounded in the concept of heritage

science as an applied field. The inquiry also acknowledged that "basic and applied research…are inextricably intertwined" (ibid, p. 24). This lack of clarity on the nature of scientific activity within the remit of heritage science is a gap in the philosophical and theoretical canon of the field with implications for heritage data science: how can an applied field be exploratory?

Both basic and applied research contribute to the aims of heritage science. Basic, or 'pure' research is "dedicated to managing and increasing knowledge of general validity" (Roll-Hansen, 2009, p. 3). Roll-Hansen distinguishes applied science as the area of intersection between science and politics: "It depends highly on advanced scientific knowledge and methods but is dedicated to the solution of practical economic, social and political problems rather than the further development of such knowledge and methods" (ibid). The distinction between basic and applied research has been summarised as a dichotomy between fact and value (Proctor, 1991). Value is at the core of understanding materials and change within heritage science (Douglas-Jones, et al., 2016). This would seem to suggest that heritage science should be classified as an applied science. However, a more enlightened perspective provided by Putnam appreciates the *distinction*, or 'entanglement' between fact and value, rather than a strict dichotomy (2002). Through the lens of current approaches within heritage science, this is demonstrated by damage functions, in which the value component is often decoupled from the dose-response function (Strlič et al., 2013). This demonstrates the integrated role of basic and applied approaches within heritage science, supporting their inclusion within its remit.

Placing data-driven approaches at the core of heritage science strengthens the argument for heritage data science to be, at least in part, a fundamental science. However, this discrepancy demonstrates the limited capacity in classifying heritage research as either basis or applied.

Models for classifying and organising scientific activity

Several models attempt to capture research activity more holistically than a dichotomy between basic and applied research which may be applicable for heritage data science.

The New Production of Knowledge proposes two ways of undertaking research which each correspond to a different kind of knowledge: Mode 1 represents the traditional academic and discipline-oriented research and knowledge; Mode 2 "operates within the context of application" and is "transdisciplinary rather than mono- or multidisciplinary" (Gibbons et al., 1994, p. vii). This latter aspect resonates with the ideology and practice of integrating several disciplines to address heritage science challenges. Similarly, Mode 2 is organisationally transient as it does not have a stable hierarchy and is more reflexive and accountable to society. Mode 2, 'the new production of scientific knowledge' thus erases the distinction between basic and applied research.

The Quadrant Model of Scientific Research was proposed by Stokes (2011). Basic and applied research are presented in the context of considerations of use and the quest for fundamental understanding. Using Pasteur's work in microbiology as an example, Stokes describes it as 'use-inspired' (or 'purpose-driven') basic research, demonstrating that basic research can be designed and undertaken in consideration of use. It is therefore 'applied', as it is informed by social and economic drivers, and 'basic', since it contributes more broadly to scientific understanding. In contrast, the work of Bohr and Edison, whose primarily contributions were in understanding atomic structure and quantum theory and developing devices for power generation and communication. 'Pasteur's Quadrant' demonstrates that interactions between theoretical and practical problems can be highly productive (Roll-Hansen, 2009). Consideration of use is an important element of heritage science research; within the Quadrant Model use-inspired basic research and applied research can thus be considered valid

components of heritage science. Additionally, as identified by the House of Lords inquiry, the complexity of "basic science underpinning conservation" (House of Lords Science and Technology Committee, 2006, p. 80) is equally important, and thus should be included within the remit of heritage data science.

Dudley (2013) proposed a three pillars model, in which basic research, use-inspired research, and development and industry share boundaries. The model includes a funding axis, since the question of the amount of funding allocated to each area is unavoidable. This model was developed based on concerns "that the quadrant model minimizes the interface between fundamental research and industrial development, giving the misleading impression that research performed in Pasteur's quadrant has the greatest impact on industry. This erroneous impression has given rise to the paradigm of use-inspired research that dominates current thinking." (ibid, p. 339). This model is similar to classifications within research and development (R&D) used by the OECD (Frascati, 2015): basic research, applied research, and experimental development. Experimental development is defined as "systematic work, drawing on existing knowledge gained from research and practical experience and producing additional knowledge, which is directed to producing new products or processes or to existing products or processes" (ibid, p. 45).

A model for heritage data science

Models that separate basic and applied research as distinct entities are not appropriate within a heritage science context, since a dichotomous model cannot capture the complex and diverse nature of scientific activities relevant to heritage data science. Although the OECD and Dudley models also acknowledge the role of industry and development within the research landscape, they both depend on distinguishing between the arbitrary distinction between applied and basic research.

Heritage data science should adopt the Quadrant Model within the context of Mode 2 as set out in *The New Production of Scientific Knowledge*. This acknowledges that heritage data science can be basic (fundamental or purpose-motivated) and applied, operates within the context of application, and involves several disciplines. The inclusion of use-inspired basic research emphasises that heritage science can produce new knowledge that is relevant for the wider scientific community. This contextualised model could be further developed by introducing a third dimension: impact potential. Rather than a binary classification of 'yes' or 'no', a continuous relative scale might be more suited. While impact is notoriously difficult to measure (and quantify) in science at large (Ravenscroft et al., 2017) it is specifically challenging within the heritage domain (Dillon et al., 2014; Katrakazis et al., 2018). This model would acknowledge the diverse range of potential impacts and their respective timelines. For example:

- Basic research might reveal a previously unknown and imminent threat that would then dictate immediate research priorities in research areas with consideration of use;
- Use-inspired basic research could demonstrate the potential of a novel technology to be applied within a heritage context, although it might not be developed commercially (or become commercially viable) for several years;
- Applied research might produce an innovative management framework that addresses a heritage-specific need; this could be implemented within a heritage organisation in a relatively short time frame.

These examples (and subsequent extensions) demonstrate an important aspect of this model: different kinds of research activity can inform the others. The unexpected imminent threat feeds into subsequent use-inspired basic research and applied research. A novel technique developed might then require applied research to develop it and hone its operation and design to heritage applications. Adapting the Quadrant model to heritage science, while including a more diverse

range of scientific activities within its remit, encourages further interaction between the modes of research. Acknowledging the contribution of basic research within heritage data science enforces its transdisciplinary identity and reinforces its exploratory nature.

Opportunities and challenges

There are innumerable active areas of research and application within heritage data science: far too many to discuss in-depth or even list herein. Some key emerging areas include climate resilience, aggregated scientific analysis, critical heritage studies, as digital documentation, heritage in crisis, digital heritage, open and linked data, and citizen science and crowd-sourced approaches. Across these areas, opportunities are enabled by key developments in areas such as open and linked data, data standards, citizen and crowd-sourced science, critical communication, and Bayesian and fuzzy approaches. These areas are active within both research and practice and rapidly developing in capability and scope.

While these emerging themes within data science holds the promise of transforming the heritage sector, there remain several challenges to its successful implementation. Some of these are challenges are universal to data science, while others are specific to the characteristics of the heritage sector.

Data quality

The colloquial phrase about data quality is ‘garbage in – garbage out’, acknowledging that poor quality data leads to unreliable output. To be of good quality, data needs to be fit for purpose within its intended use(s), such as decision making and planning (Redman, 2008); high-quality is a true and accurate representation of the real-world entity it should represent with limited bias. These present particular challenges for heritage data.

Heritage data, especially data *about* heritage, may have implicit bias or have caveats with implications for outputs. Heritage data science draws on a diverse range of data sources: many of which exist to produce data for their own purposes, for which the heritage application(s) are secondary. For example, heritage data science makes frequent use of climate data typically collected for regional and large-scale meteorological and climate monitoring (e.g., Orr et al., 2018; Brimblecombe et al., 2020). One of the challenges in using these datasets within a heritage context is determining if they are suitably representative of the localised environment relevant to a heritage context. Similarly, this data is often collected at time intervals that do not represent the timescales of heritage phenomena. With a move toward open and FAIR data within and beyond the heritage sector (data that is findable, accessible, interoperable, and reusable), these types of challenges will become more prevalent. Thus, heritage data scientists must determine whether this data is sufficiently accurate and precise for the task, or whether it is necessary to produce a dataset with suitable spatial and temporal coverage.

The landscape of heritage data can be discriminatory. As chronicled in academic literature (Hoffmann, 2019) as well as popular reading (O'Neil, 2016), algorithmic decision making implemented with the best intentions can be discriminatory. These often incorporate proxy data: data that is accessible or seemingly insensitive that indirectly represents an important factor within the context as a substitute for data that is difficult to collect or sensitive. These data may be correlated with other factors such as race, gender, and sexuality, or other sensitive characteristics, many of which have legislative protection in several regions.

Heritage data can be rooted in legacies that do not represent contemporary perspectives and discourse. A significant amount of heritage data has been curated by large cultural organisations or so-called 'memory institutions' that aim to preserve, contextualise, and communicate canonical elements of culture, historical narrative, and collective memory. These institutions, and the social memory they have produced, are to varying extents rooted in West

centrism, colonialism, and the world views of the social and moral values of the upper-class elite at the helm of their establishment and operation (West, 2010; Smith & Waterton, 2012). This manifests in the curation, management, and dissemination of heritage, which is increasingly embodied in the relevant heritage data. Thus, heritage data may represent particular narratives and themes, downplaying the importance of diverse and often conflicting social histories and perspectives. The implication for heritage data science is that any work undertaken with this data must contextualise output within this context, and strive to supplement both data and discussion accordingly. Citizen and crowd-sourced methods holds particular promise to supplement existing mainstream narratives and generate data: a successful example is *Pride of Place: England's LGBTQ Heritage* (Historic England, 2020). This initiative acknowledged that LGBTQ histories are embedded in England's built, cultural, and natural landscapes, but was severely underrepresented in heritage documentation. The initiative uses a map-based crowd-sourced approach to produce a geolocated dataset representing LGBTQ stories and places. In the absence of suitable supplementary data sources, heritage data science projects should transparently discuss potential bias and demonstrate an understanding of its culturally-embedded nature.

Data integrity

Data integrity ensures consistency and accuracy of data over its entire lifetime, which is crucial to the successful implementation of data-driven decision making. The challenges posed to data integrity within heritage data science are related to both physical assets, digital assets, and assets that have both physical and digital representations.

Physical heritage assets are an important source of heritage data. Beyond their metadata, they can provide new insight when made the subject of scientific investigation or perspective surveys. However, these data are fixed in time at their instance of collection, representing a

snapshot. In reality, physical heritage assets are dynamic and ever-changing, as are perceptions of their physicality and value. Thus, data representing the state of a heritage collection may accurately represent its conditions, but this may be invalidated by subsequent changes of assets. Heritage data science needs to demonstrate awareness of the static nature of many heritage data, and develop flexible approaches that respond to physical change and collection demographics, as well as emerging expectations of physical heritage assets and attitudes toward heritage.

Born digital assets equally undergo inevitable change, but their formats pose more significant challenges to long-term sustainability. Prone to natural bit-rot (a digital analogue of physical material decay; see Cerf, 2011), digital materials require regular checks and maintenance to be implemented as part of their preservation. The interfaces used to interrogate heritage data change; thus, data that was retrievable and useful at the time of acquisition may prove challenging to access as platforms and software become obsolete. Therefore, programmes of conversion and updating of data must be embedded into the long-term management of born digital assets to ensure they can be retrieved and accessed. For example, archived materials from early pioneers who incorporated CAD (computer aided design) and 3D software into architectural practice pose challenges. While the bespoke files hold a rich array of information pertaining to design process, curation and conversion may result in these becoming inaccessible. The remaining material after conversion may simply be a 2D rendering of the models, or presentation material relating to the models. However, how to reconcile conversion and maintenance with the potential to lose heritage value remains an open challenge within heritage data science.

By nature, sustainable heritage must consider long time horizons, often seeking data that can adequately represent the dynamic nature of heritage and its context over them. Within a data science paradigm, most *forward* considerations (e.g., forecasting and modelling) are born-

digital. However, there are a significant number of records reflecting the past that require digitisation in order to be incorporated into data science approaches. One example is the 'Data Rescue initiative as part of the ACRE project run by the UK Met Office (http://www.met-acre.net). This initiative seeks to develop a comprehensive model of historical climate back to the early 19th century based on paper-based archives, such as the diaries of explorers and 'gentleman' scientists, reports, private archives, the records of port authorities and ship logs books. A significant amount of resources are required to identify, digitise, and process these types of data, although natural language processing and citizen science can make important contributions.

More broadly, digitised assets may not be as rich as physical assets. Although they may include important information including essential metadata and text, contextual information may be lost. For example, watermarks have been shown to be an essential component of historical investigation of early printed documents (Calì, 2018). Depending on the digitisation procedure and storage format, important characteristics of the asset may not be included in a digitised record.

Transdisciplinarity and value-creating science

Heritage data science is by its nature culturally embedded. Resultantly, those involved (researchers, partner organisations, the public as a stakeholder) may hold vastly different views on how a heritage data science project should be undertaken. From the outset of a heritage data science project, care should be taken to identify the diverse array of stakeholders. Input should be solicited throughout the project's lifetime, from conceptualisation, acquisition, analysis, to dissemination and long-term curation.

By nature of studying heritage, heritage data science produces and enriches heritage value. Therefore, it is paramount to demonstrate awareness of what might be considered a

perpetuating cycle of data science. Generating data and developing data-driven insights from it facilitates opportunities for further heritage data science within the area. Thus, heritage data science, and the topics it addresses, risk perpetuating a positive-reinforced feedback cycle in which heritage typologies and challenges studied become further studied, at the risk of others being neglected. Heritage data science should seek to identify cross-cutting expertise and methods that hold promise to address several heritage typologies and incorporate several overlapping heritage value frameworks and stakeholder perspectives.

Digital literacy

As a transdisciplinary field, heritage data science requires equal command of data and data-driven methods, the heritage domain, and imaginative thinking. This contrasts a multidisciplinary of interdisciplinary mode of working, in which those with expertise within the heritage domain work alongside those with relevant data skills. This poses a significant challenge to the training of heritage data scientists and upskilling of those working in the heritage sector.

Specialised training is required to produce transdisciplinary heritage data scientists who are equipped to address sustainability challenges within the heritage sector. Until recently, very few academic courses incorporated elements of heritage science data analysis, visualisation, use and reuse, digitisation, and data science (Albuerne et al. 2018). However, university-level courses are increasingly combining a deep understanding of the heritage domain with state-of-the-art computer science. One recent initiative is the MSc *Data Science for Cultural Heritage* at UCL, https://www.ucl.ac.uk/bartlett/heritage/study/data-science-cultural-heritage-msc), that address this gap by producing researchers and data scientists who are equipped to understand the complexities of working in the heritage domain, as well as more broadly in other challenging domains. Similarly, programmes focusing on conservation and restoration and

heritage are increasingly embedding digital and statistical skills into their training, such as a two-week intensive module within the MSc in Conservation and Restoration of Cultural Heritage at the University of Amsterdam and plans to shortly integrate data science into the curriculum in conservation-restoration at the University of Antwerp.

More broadly, the heritage sector is realising that the future of work is going to be increasingly data driven. Initiatives such as the *Mapping the Museum Digital Skills Ecosystem* project (Barnes et al., 2018) and the Heritage Alliance's *Heritage Digital* programme identify gaps and targeted training to upskill the heritage workforce. There is an ongoing need for flexible and accessible training that recognises the unique challenges posed to data science within the heritage sector.

Conclusions

Heritage data science is an emerging transdisciplinary field that informs heritage engagement with heritage and its interpretation and long-term management using data-driven approaches. It is also highly innovative and produces new knowledge and furthers understanding of society and the environment. Heritage data science requires in-depth comprehension of the complexities of the heritage domain and an awareness of science as culturally embedded enquiry, expertise in all aspects of the data pipeline, and an appreciation for the role of human intelligence and critical reasoning in research and decision making. There are several open challenges within heritage data science: data quality and integrity especially with a focus on equality, diversity, and inclusion; a lack of frameworks for enabling transdisciplinary data-driven science in a heritage context; and, limited educational pathways for both educating data scientists to work within the heritage domain and upskilling heritage professionals in preparation for the imminent digital transformation within the sector.

References

Albuerne, A., Grau-Bove, J., & Strlic, M. (2018). The Role of Heritage Data Science in Digital Heritage. In *EuroMed 2018: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection* (pp.616-622). Cham: Springer. doi: 10.1007/978-3-030-01762-0_54

Barnes, S. A., Kispeter, E., Eikhof, D. R., & Parry, R. (2018). *Mapping the Museum Digital Skills Ecosystem - Phase One Report*. Leicester: University of Leicester. doi:10.29311/2018.01

Brimblecombe, P., Hayashi, M., & Futagami, Y. (2020). Mapping Climate Change, Natural Hazards and Tokyo's Built Heritage. *Atmosphere*, *11*(7), 680. doi:10.3390/atmos11070680

Calì, C. (2018). The importance of a project to enhance the watermarks of the Codex Atlanticus by Leonardo Da Vinci. In *5th International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2018* (pp.333-338). Retrieved from https://re.public.polimi.it/handle/11311/1116346

Cao, L. (2017). Data science: challenges and directions. *Communications of the ACM*, *60*(8), 59-68. doi:10.1145/3015456

Cao, L. (2018). Data Science Thinking. In *Data Science Thinking* (pp. 59-90). Cham: Springer. doi:10.1007/978-3-319-95092-1_3

Cerf, V.G. (2011). Avoiding "bit rot": Long-term preservation of digital information [point of view]. *Proceedings of the IEEE*, *99*(6), 915-916. doi:10.1109/JPROC.2011.2124190

Dillon, C., Bell, N., Fouseki, K., Laurenson, P., Thompson, A., & Strlič, M. (2014). Mind the gap: rigour and relevance in collaborative heritage science research. *Heritage Science*, *2*(1), 11. doi:10.1186/2050-7445-2-11

Douglas-Jones, R., Hughes, J.J., Jones, S., & Yarrow, T. (2016). Science, value and material decay in the conservation of historic environments. *Journal of Cultural Heritage*, *21*, 823-833. doi: 10.1016/j.culher.2016.03.007

Dudley, J.M. (2013). Defending basic research. *Nature Photonics*, *7*(5), 338-339. doi: 10.1038/nphoton.2013.105

Eklund, L., Sjöblom, B., & Prax, P. (2019). Lost in Translation: Video Games Becoming Cultural Heritage?. *Cultural Sociology*, *13*(4), 444-460. doi:10.1177/1749975519852501

Frascati, M. (2015). Guidelines for Collecting and Reporting Data on Research and Experimental Development, 7th ed. Paris: OECD Publishing.

Fredheim, L.H., & Khalaf, M. (2016). The significance of values: heritage value typologies re-examined. *International Journal of Heritage Studies*, *22*(6), 466-481. doi:10.1080/13527258.2016.1171247

Gibbons, M. (Ed.) (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. Atlanta, GA, USA: Sage.

Gray, J. (2004). Online Science: the 20 questions approach. *Scientific Data Symposium*, Redmond, WA, 25 May 2004. Retrieved from http://jimgray.azurewebsites.net/talks/SciData.ppt

Hey, T. (Ed.) (2009). *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Historic England (2020). *Pride of Place: England's LGBTQ Heritage.* Accessed 24 August, 2020. Retrieved from https://historicengland.org.uk/research/inclusive-heritage/lgbtq-heritage-project/

Hoffmann, A.L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900-915. doi:10.1080/1369118X.2019.1573912

House of Lords Science and Technology Committee (2006). 9th Report of Session 2005–06: Science and Heritage, Report with Evidence. HL Paper 256.

Katrakazis, T., Heritage, A., Dillon, C., Juvan, P., & Golfomitsou, S. (2018). Enhancing Research Impact in Heritage Conservation. *Studies in Conservation*, *63*(8), 450-465. doi:10.1080/00393630.2018.1491719

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Atlanta, GA, USA: Sage.

NHSF (2018). Strategic Framework for Heritage Science in the UK, 2018-2023. National Heritage Science Forum. Retrieved from http://www.heritagescienceforum.org.uk/what-we-do/strategic-framework

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY, USA: Crown Publishing Group.

Orr, S.A., Young, M., Stelfox, D., Curran, J., & Viles, H. (2018). Wind-driven rain and future risk to built heritage in the United Kingdom: Novel metrics for characterising rain spells. *Science of the Total Environment*, *640*, 1098-1111. doi:10.1016/j.scitotenv.2018.05.354

Palfrey, J.G., & Gasser, U. (2011). *Born digital: Understanding the first generation of digital natives*. New York, NY: Basic Books.

Proctor, R. (1991). *Value-free science?: Purity and power in modern knowledge*. Cambridge, MA, USA: Harvard University Press.

Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA, USA: Harvard University Press.

Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PloS one*, *12*(3), e0173152. doi:10.1371/journal.pone.0173152

Redman, T.C. (2008). *Data driven: profiting from your most important business asset*. Cambridge, MA, USA: Harvard Business Press.

Roll-Hansen, N. (2009). *Why the distinction between basic (theoretical) and applied (practical) research is important in the politics of science*. London School of Economics and Political Science, Contingency and Dissent in Science Project.

Schreibman, S., Siemens, R., & Unsworth, J. eds., 2008. *A companion to digital humanities*. Hoboken, NJ, USA: John Wiley & Sons.

Smith, L., & Waterton, E. (2012). Constrained by commonsense: The authorized heritage discourse in contemporary debates. In R. Skeates, C. McDavid, & J. Carman (Eds.) *The Oxford Handbook of Public Archaeology* (pp.153-171). Oxford, UK: Oxford University Press.

Stokes, D.E. (2011). *Pasteur's quadrant: Basic science and technological innovation*. Wasington, DC, USA: Brookings Institution Press.

Strlič, M., Thickett, D., Taylor, J., & Cassar, M. (2013). Damage functions in heritage science. *Studies in Conservation*, *58*(2), 80-87. doi:10.1179/2047058412Y.0000000073

Terras, M. (2011). Quantifying Digital Humanities (PDF). *UCL Centre for Digital Humanities*. Retrieved 24 August, 2020. Retrieved from http://www.ucl.ac.uk/infostudies/melissa-terras/DigitalHumanitiesInfographic.pdf

Tress, G., Tress, B., & Fry, G. (2007). Analysis of the barriers to integration in landscape research projects. *Land use policy*, *24*(2), 374-385. doi:10.1016/J.LANDUSEPOL.2006.05.001

Tukey, J.W. (1962). The future of data analysis. *The annals of mathematical statistics*, *33*(1), 1-67. doi:10.1214/aoms/1177704711

UNESCO (2003). Charter on the Preservation of Digital Heritage. Retrieved from http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html

Vecco, M. (2010). A definition of cultural heritage: From the tangible to the intangible. *Journal of Cultural Heritage*, *11*(3), 321-324. doi:10.1016/j.culher.2010.01.006

West, S. (2010). Heritage and class. In: R. Harrison (Ed.) *Understanding the Politics of Heritage. Understanding Global Heritage* (pp.270–303). Manchester and Milton Keynes, UK: Manchester University Press and The Open University.