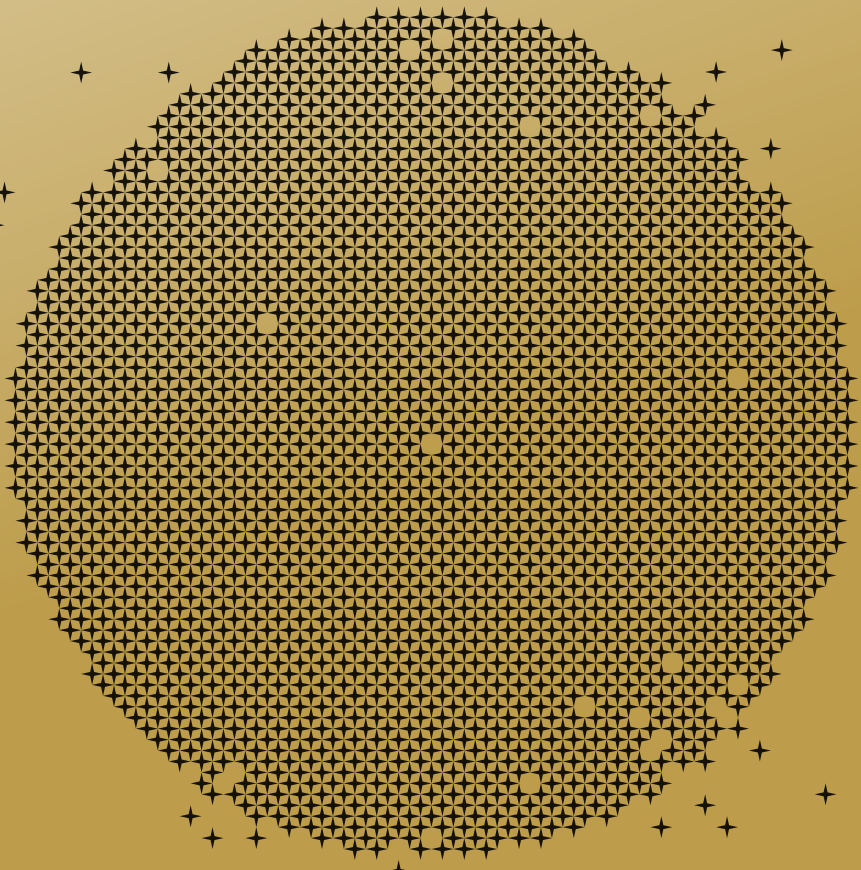
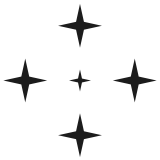
DISEÑO

VII/01
journal of design culture
_Designing Digital Humanities





Disegno

Journal of Design Culture
Double-blind peer-reviewed, open access scholarly journal

Aims and Scope

Disegno publishes original research papers, essays, and reviews on all aspects of design cultures. We understand the notion of design culture as resolutely broad: our aim is to freely discuss the designed environment as mutually intertwined strands of sociocultural products, practices, and discourses. This attitude traverses the disciplinary boundaries between art, design, and visual culture and is therefore open to all themes related to sociocultural creativity and innovation. Our post-disciplinary endeavour welcomes intellectual contributions from all members of different design cultures. Besides providing a lively platform for debating issues of design culture, our specific aim is to consolidate and enhance the emerging field of design culture studies in the Central European academia by providing criticism of fundamental biases and misleading cultural imprinting with respect to the field of design.

All research papers published in Disegno undergo a rigorous double-blind peer review process.
This journal does not charge APCs or submission charges.

Contact: *Moholy-Nagy University of Art and Design*
H-1121 Budapest, Zugligeti út 9–25.
disegno@mome.hu

The full content of Disegno can be accessed online: ***disegno.mome.hu***

Published by: *József Fülöp*
Publisher: Moholy-Nagy University of Art and Design, 1121 Budapest, Zugligeti út 9-25.

ISSN: *2064-7778 (print)* ***ISSN:*** *2416-156X (online)*

MOHOLY-NAGY
művészeti egyetem
university of art and
design budapest

DISEGNO_VII/01_DESIGNING DIGITAL HUMANITIES

Contents

DATA AND KNOWLEDGE MODELLING AS THE METHODOLOGICAL FOUNDATION OF DIGITAL HUMANITIES

Dinara Gagarina

ABSTRACT

Digital humanities is a multidisciplinary field that leverages digital technology and methodologies to explore and answer questions pertaining to the humanities. It is a dynamic intersection between the domains of computer science and the humanities, promoting innovation, collaboration, and research at the highest levels. However, as a relatively young field, the methodological foundations of the digital humanities are still being established. This paper seeks to explore the core methodologies that underpin digital humanities.

The modelling of data, information, and knowledge can be considered one of the foundations of digital humanities. One of the arguments confirming this is that the development of digital humanities and the development of technologies in general are the development of ways to formalise and present data and knowledge. Science has come a long way from the modelling and computer representation of numbers to generating texts and art on the basis of prescribed inputs. With the advent of artificial intelligence, especially machine learning and deep learning techniques, the potential for more sophisticated and nuanced data modelling in the digital humanities has expanded significantly, linking computational capabilities with humanistic inquiries in unprecedented ways.

The article considers the periodization, classification, and trends of approaches and methods for modelling data, information and knowledge in the humanities. The article provides an overview of existing examples and data models of different complexity from various humanities disciplines, including history, linguistics, literary criticism, and cultural studies.

#data modelling, #knowledge representation, #semantic networks, #GPT, #digital history

https://doi.org/10.21096/disegno_2023_1dg

INTRODUCTION

The transformation of the humanities through the adoption of digital technologies has led to a new age of scholarship. The rapidly developing field of digital humanities combines traditional humanities disciplines such as history, literature, philosophy, and art with computer science, focusing on the use of computational tools for the analysis, visualization, and understanding of human culture. This confluence of digital tools and humanistic enquiry is reshaping how we understand and represent the vast spectrum of human experience, particularly in an era of big data and increasing digitisation of all kinds of historical and cultural heritage.

The term "Digital Humanities" is an umbrella term that covers a wide range of activities and disciplines (Gold 2012; Burdick et al. 2012). Because of its integrative and interdisciplinary nature, the field of digital humanities is constantly exploring its methodological core and establishing methodological foundations that can guide its research, applied projects, and educational endeavours. Many scholars have written about the need for a methodological core or foundation in digital humanities, as it is a vital part of establishing the field's credibility and facilitating its further development. Two notable figures in this conversation are Patrik Svensson and Paul Rosenbloom. Svensson's work deals with the humanities computing as a field, considers how the digital humanities could be considered a field or a discipline, and discusses methodological commonplaces (Svensson 2010). Rosenbloom's work, on the other hand, engages more deeply with the computational aspects of the digital humanities, providing a framework for building a multi-level methodology for the digital humanities (Rosenbloom 2012).

As one of the fundamental pillars, data and knowledge modelling plays a pivotal role in shaping the future of digital humanities. Data and knowledge modelling refers to the process of creating structured representations of information. These models offer a framework for organising, categorising, and analysing data. In the context of the digital humanities, data and knowledge modelling allows for the representation of complex concepts, narratives, and relationships in a form amenable to computational analysis. The assertion that

modelling forms the bedrock of the digital humanities has gained wide acceptance (McCarty 2005, Flanders and Jannidis 2015). However, it is still a complicated field with many attempts to develop this insight further and conceptualize the modelling and its associated characteristics in the digital humanities (Jannidis 2018; Ciula et al. 2018).

Recent advancements, including the rise of large language models like GPT-3 and the broad adoption of AI techniques, have driven significant growth in the digital humanities. This growth underscores the urgent need to revisit the foundational methodological pillars of digital humanities research and its applications. Central to this discourse is data and knowledge modelling, a crucial methodology in the field. This approach allows for the computational analysis of cultural artifacts and dynamics at a large scale but also runs the risk of oversimplifying intricate humanistic nuances. Such analyses, especially when incorporating state-of-the-art tools like BERT for computational text analysis, require careful human interpretation and judgment. Moreover, the increasing influence of large language models is reshaping knowledge representation within the digital humanities. This paper aims to critically evaluate these methodologies, emphasising both their capabilities and limitations, especially amidst the rapid technological advancements. It posits that while these techniques offer a robust framework for enquiry, they mandate ethical application and critical evaluation, a sentiment growing in importance due to the swift progression of AI. Furthermore, as AI continues to evolve, fostering discussions both within and outside the digital humanities community about its implications becomes essential, ensuring that emerging tools align with humanistic values.

THE HUMANITIES AND EVOLUTION OF DATA AND KNOWLEDGE MODELLING

The progression of data and knowledge modelling has shown significant development over time. Primitive models were often rudimentary, linear, and devoid of the capacity to represent intricate relationships. In contrast, contemporary models are proficient in demonstrating a diverse array of relationships, attributes, and entities. Several mathematical, computational, and philosophical pioneers, such as Claude Shannon, Alan Turing, John von Neumann, Donald Knuth, Marvin Minsky, and Judea Pearl, have been formulating theoretical models of computers and computing for approximately a century (Shannon 1948; Turing 1937; von Neumann 1945; Knuth 1968; Minsky 1986; Pearl 2000). Simultaneously, a shift has occurred in the comprehension of modelling and knowledge formalisation within the realm of the humanities.

In the formation of theoretical and conceptual foundations for modelling within the humanities, numerous influential figures have emerged, including Johanna Drucker, Willard McCarty, Lev Manovich,

Ted Underwood, Manfred Thaller, Peter Doorn, among others (Drucker 2014; Drucker 2021; Fiormonte et al. 2015; McCarty 2005; McCarty 2018; Manovich 2013; Underwood 2019; Thaller 1985; Thaller 2012; Doorn 2021). For instance, in "Graphesis: Visual Forms of Knowledge Production", Johanna Drucker (2014) emphasized the significance of graphical systems in the knowledge production process, thereby addressing the challenges and opportunities of modelling humanities data. Similarly, in "Humanities Computing" Willard McCarty (2005) underscored the methodological implications of data modelling within the humanities, providing a theoretical basis for the discipline of the digital humanities.

Noteworthy research has been conducted by pioneers in the digital humanities field, spanning the entire scope of the humanities as well as specific disciplines such as history, literary criticism, and cultural studies. Lev Manovich's "cultural analytics" demonstrate the potential of data modelling in visual culture and art history. Ted Underwood's work showcases the efficacy of data modelling in literature. Manfred Thaller and Peter Doorn have significantly contributed to the modelling of historical databases and information systems (Doorn 2021; Thaller 1985; Thaller 2012). Each of these scholars, with their distinct areas of focus and methodologies, has significantly contributed to the progression of data and knowledge modelling in the humanities.

We can broadly segregate the evolution of the formalisation and modelling of knowledge within the humanities into four distinct stages.

PRE-DIGITAL AND EARLY DIGITAL PHASES

In the pre-digital and early digital phases, scholarly works in the humanities were fundamentally analogue, leaning heavily on written texts, artefacts, or oral traditions. Information was typically organised in the form of documents, with knowledge predominantly catalogued through indices. During this phase, data was primarily unstructured, preserved as manuscripts, books, and other physical documents, with knowledge representation being largely narrative and qualitative.

Notable instances from this pre-digital era abound across various humanities domains. One significant example includes the development of extensive catalogues and indices designed to aid scholars in locating works within vast collections comprising books, maps, manuscripts, and more. Additionally, we can reference various historical atlases seeking to represent data visually and spatially.

A seminal early example of digital humanities is the Index Thomisticus project spearheaded by Roberto Busa in the mid-20th century. This project, which entailed creating a comprehensive index and concordance of the works of Thomas Aquinas and related authors, initially took form as a print project before transitioning into a digital medium, thus representing one of the earliest instances of digital humanities (Busa 1980).

In this phase, we can also discern preliminary endeavours in data modelling, chiefly stemming from early attempts at representing numerical data via basic computational techniques. Here we can cite the emergence of statistical analyses in disciplines like history and sociology, as well as the rise of economic history and historical demography as initial examples of digital or mathematical history. A notable figure here is Robert Fogel, who shared the 1993 Nobel Prize in Economics with Douglass North. Fogel used innovative statistical techniques to examine the economic impact and efficiency of institutions such as railways and slavery, as discussed in his significant work "Time on the Cross: The Economics of American Negro Slavery" (Fogel and Engerman 1974), which marked a pivotal point in economic history. Despite inciting controversy, Fogel's methodologies further spurred the development of economic analysis in historical processes.

DIGITAL PHASE

The digital phase in humanities commenced with the advent of digital computers, which enabled the digitisation, storage, and manipulation of enormous volumes of data. Fundamental data models such as flat files and relational databases gained popularity during this era. The introduction of markup languages, including XML and HTML, facilitated the encoding of semantics in texts, heralding the digital representation and organisation of information (Riley 2017).

Relational, hierarchical, and graph data models emerged as significant areas during this phase. As an example, relational databases were central to early digital humanities projects such as the Perseus Digital Library at Tufts University, which aimed to collect and provide access to classical texts and artifacts (Mylonas 1993). This project utilised relational databases to store and organise information in a structured manner, thereby paving the way for more complex digital humanities projects.

Other examples of innovative projects during the digital phase include the Text Encoding Initiative (TEI), which was founded in 1987. This initiative developed a standard for encoding machine-readable texts in the humanities and social sciences, with a specific emphasis on the markup of texts (Sperberg-McQueen and Burnard 1994). The development of graph models led to the rise of network analysis in the digital humanities. An instance of this is the work of Scott Weingart, a digital humanities specialist, who applied network analysis to historical studies, thus offering new insights into historical relationships and processes (Weingart 2011). Lastly, projects like Manfred Thaller's (1985) "Beyond Collecting: On the Design and Implementation of CLIO, a DBMS for the Historical Sciences" showcased the power of databases

and data modelling in digital humanities research. These early pioneers laid the groundwork for the sophisticated models and techniques that are now commonplace in the field.

SEMANTIC PHASE

The semantic phase of the digital humanities represents a significant advancement in data modelling, where the focus shifts towards creating intricate representations of knowledge using advanced semantic technologies and artificial intelligence. Here, it is the establishment of ontologies, semantic networks, and knowledge graphs that have led to a more nuanced understanding and depiction of the relationships and complex concepts within humanities research.

A seminal work in this phase is the project "Pelagios" led by Leif Isaksen, Elton Barker, Rainer Simon, and Pau de Soto. The project aims to create a comprehensive semantic graph of ancient places, annotated from a vast array of resources spanning multiple millennia and languages. It applies semantic web technologies to create an interconnected web of historical geographical information (Isaksen et al. 2014). In this phase, TEI begins to play more and more increasingly important role. An example of semantic modelling in digital humanities is seen in "Prosopographies" and "TEI Guidelines" which refer to the development of detailed collections of people, networks, and relationships in historical or literary contexts, usually encoded in TEI or similar XML-based schemas (McCarty 2004; Romanello et al. 2013). The TEI Guidelines enable scholars to represent complex textual phenomena and their semantics in a standardized way, which is crucial for interoperability and data exchange in the digital humanities (Sperberg-McQueen and Burnard 1994). A further pioneering project in this phase is the Linked Jazz Project, which uses linked open data technologies to uncover relationships between jazz musicians based on data extracted from interviews and other documents (Liu and Pattuelli 2013). This project exemplifies the way semantic web technologies can be applied in digital humanities research to uncover and visualize complex networks of relationships. Kim et al. (2017) investigate the relationship between literary genres and emotional plot development comparing different models that use emotion-related information to classify genres of stories with traditional bag-of-words models for genre classification. They find that different genres have different emotional arcs, with some genres showing more uniform emotional development than others. In digital history, knowledge graphs and semantic web technologies are uncovering new connections between historical events and figures (Meroño-Peñuela et al. 2014). Ontologies and linked open data help model the complexity of historical relationships (Ide, 2007).

Recent years have seen remarkable advances in data and knowledge modelling techniques across various humanities domains. In literary analysis, models like BERT (Bidirectional Encoder Representations from Transformers) are revolutionising computational text mining (Devlin, 2018). Luccioni and Rogers (2023) discuss the use of Large Language Models (LLMs) in Natural Language Processing (NLP) research, highlighting the limitations and challenges associated with their evaluation and impact on the field. They emphasize the need for rigor, transparency, and diversity in research approaches, as well as reproducibility and access to resources.

The semantic phase revolutionizes how we understand and interconnect data in the digital humanities. It has allowed for the digital representation and analysis of complex relationships in the humanities in a way that was not possible with previous approaches.

GENERATIVE ARTIFICIAL INTELLIGENCE PHASE

The utilisation of generative AI and its associated models in the humanities presents revolutionary avenues for exploration and interpretation. Generative AI models, such as OpenAI's Generative Pretrained Transformer (GPT), leverage vast datasets to produce novel content that mirrors learned patterns and structures.

The application of generative artificial intelligence and associated models within the humanities heralds innovative and thrilling possibilities for research and interpretation, potentially deepening and refining our understanding of human culture and history. In literary studies, these models hold potential for style analysis, thematic pattern recognition, and new text generation based on training data. Models can be trained on texts from a specific author or literary period and subsequently generate content that emulates the style and theme of the source material (Jockers 2013). Historical research can also benefit from generative AI models. They can be employed to analyse and reproduce historical documents and artifacts. For instance, training these models on images of historical documents allows them to generate new images mirroring the style and characteristics of the original sources, offering an effective method for studying and recreating historical contexts and cultures. Cultural studies, too, can leverage generative AI for analysing and recreating artistic styles and cultural trends. By training models on images of art pieces, new images that emulate the style and characteristics of the training data can be generated, proving beneficial for researching and analysing artistic styles and cultural trends (Elgammal et al. 2017).

Recent large language models like GPT-4 and Claude from Anthropic promise to transform knowledge representation and text generation within digital humanities domains (Bommasani et al. 2021; Anthropic

2023). By learning patterns from vast datasets, these models can generate synthetic content mimicking various writing styles. Such capabilities allow examining textual attributes at scale, like analysing genre conventions or authorial voice across corpora.

DIGITAL HUMANITIES: METHODOLOGICAL FOUNDATIONS AND TASKS

Data and knowledge modelling can be viewed as the fundamental methodological basis of the digital humanities, as they enable the creation and implementation of computational tools and techniques in humanities research (Borgman, 2009). At the heart of these operations is the transformation of humanistic data into a computational format. This transformation encompasses two main stages: data modelling, referring to the conversion of humanities data into a computer-processable format, and knowledge modelling, which involves organising humanities knowledge into structured formats, typically comprising relationships between entities (Kitchin and McArdle 2016). These processes underpin a myriad of applications in digital humanities, from analysing linguistic patterns in literature to charting historical events and tracking cultural trends (McCarty, 2005; Rockwell and Sinclair 2016).

As the digital humanities evolve and flourish, the role of data and knowledge modelling becomes increasingly vital for several reasons.

First, data and knowledge modelling significantly foster interdisciplinary research within the digital humanities. By enabling diverse data source integration, these models facilitate a comprehensive and holistic approach to complex problems. This integration permits exploration at the intersection of various fields, thus expanding the scope of enquiry and potential discoveries.

Second, data and knowledge modelling contribute immensely to the preservation and digitisation of cultural artifacts, thus playing a crucial role in safeguarding our cultural heritage. By employing these techniques, precious artifacts are preserved and made accessible for future generations, ensuring the continuous availability of cultural resources for ongoing and future research.

Third, data and knowledge modelling enhance data analysis, a critical advantage in digital humanities. Through these techniques, scholars can perform comprehensive and scalable data analyses, a feat unachievable by traditional, non-digital methods. These advanced analyses yield novel insights and discoveries, making data and knowledge modelling powerful tools for unearthing new knowledge in the humanities (Manovich 2013).

In essence, data modelling serves as a foundation for the digital humanities. It involves converting traditional humanities sources

into machine-readable formats, such as texts, databases, or GIS files. During this process, information from primary and secondary sources is systematically encoded, annotated, and categorized, enabling computational tools to process and analyse it. This methodology facilitates the discovery of patterns, correlations, and other significant phenomena that may not be discernible through conventional humanities research methods (Drucker 2014).

The tools of data modelling play crucial roles in digital humanities, including computational analysis, data visualisation, digital archiving and preservation, and fostering a collaborative and open-source approach (Jockers 2013; Manovich 2013; Unsworth 2000; Spiro 2012):

1. Computational analysis: Using computer algorithms, researchers can manipulate and analyse large datasets, perform intricate statistical analyses, and apply machine learning techniques to identify patterns and derive insights. Text analysis, topic modelling, sentiment analysis, and network analysis are among the key computational methods employed in digital humanities.

2. Data visualisation: Visualisation plays an essential role in the digital humanities. Given the human eye's natural inclination towards and swift processing of visual representations, digital humanities scholars utilize data visualisation techniques to depict their computational analysis findings. These visualisations can range from graphs, charts, and maps to more sophisticated interactive interfaces, allowing users to examine data and findings in depth.

3. Digital archiving and preservation: Digitally preserving cultural and historical artifacts is a vital task in the field of the digital humanities. This process entails not just digitising these resources but also maintaining them in formats that ensure their longevity and accessibility for future research. Efficient data and metadata models are necessary for these tasks.

4. Collaborative and open-source approach: A foundational methodological principle in digital humanities is a commitment to collaboration and an open-source ethos. Digital humanities projects often involve teamwork, with various scholars contributing their unique skills and perspectives. Furthermore, many scholars and projects adopt an open-source approach, sharing their code, data, and findings openly with the community. This collaborative spirit promotes greater transparency, reproducibility, and knowledge sharing.

Data and knowledge modelling in the humanities is closely related to the concept of uncertainty. Edmond (2019) suggests several measures to address challenges and improve data modelling in the humanities. One approach is to focus on interoperability and comparative legibility, allowing researchers to fluidly move between different sources and perspectives. This can be achieved by enabling the combination and comparison of siloed sources, without losing

their context and complexity. Another measure is to incorporate fuzzy search capabilities that reduce false negatives and increase interrogability. By allowing for more flexible and nuanced searches, researchers can navigate the uncertainties and ambiguities inherent in humanistic data (Edmond 2019).

THE USE OF MODELLING IN THE DESIGN OF EDUCATIONAL PROGRAMS IN DIGITAL HUMANITIES

The integration of modelling in the design of educational programmes in digital humanities has substantial potential to revolutionise pedagogical approaches, engender critical thinking, and prepare students more effectively for a digital age. Scholars like Willard McCarty have advocated for the incorporation of modelling in digital humanities education, arguing that the process of constructing models can enhance students' comprehension of their subject matter (McCarty 2004).

Modelling, which serves as a cornerstone concept in digital humanities education, introduces students to the practice of representing intricate systems and phenomena in a simplified, structured manner, fundamental to much computational analysis. This notion of modelling becomes palpable when students engage with work by scholars such as Stephen Ramsay, renowned for both his use of modelling in his research and his emphasis on its role in pedagogy. His writings reflect on teaching computational literacy in the humanities, focusing on critical engagement with models (Ramsay 2011).

Moreover, the process of model construction can help students foster crucial skills such as critical thinking, problem-solving, and data literacy. This perspective is often associated with scholars like Johanna Drucker, who have contributed significantly to the pedagogical discourse in digital humanities. Drucker has extensively discussed the interpretive aspects of data modelling, and the critical use of visualisation and graphical models in humanities education (Drucker 2011).

The use of modelling also promotes a more active, project-oriented approach to learning. Instead of passively absorbing information, students engage in the creation and testing of models, which can foster a deeper understanding and engagement. This concept aligns with the contemporary pedagogical theories that emphasise the importance of active learning and real-world application of knowledge. In this context, the work of scholars like Ted Underwood is notable. Underwood is known for his application of machine learning and statistical modelling in literary studies and has actively discussed how these modelling techniques can be incorporated into digital humanities curricula (Underwood 2014).

Lastly, modelling can be applied in structuring the design of the curriculum itself. As a model can represent a complex system, it can

similarly represent the structure and sequence of an educational programme. This perspective can help educators identify gaps in the curriculum, logically sequence courses, and ensure that learning outcomes are aligned with instructional activities and assessments. Elijah Meeks, a digital humanities specialist with a strong focus on data visualisation, has written about using these methods in an educational context (Meeks 2015).

In our endeavour to develop digital humanities educational programs, we have conducted extensive experiments between 2016 and2018 that have confirmed the practical effectiveness of utilising data and data modelling as a foundation for program structure (Gagarina, Kornienko 2018). We explored a variety of strategies for structuring courses, such as aligning it with different humanities fields, digital humanities sections, methodologies, technologies, software, or data types. Across several academic years, we examined two course formation approaches: a collaborative method, which involved both students and teacher, and a data-centric method. Remarkably, we found that the data-centric approach, which built the course around different types of data and its modelling, was more effective for students, especially for those who had limited exposure to traditional humanities disciplines.

The integration of modelling in the design of educational programmes in digital humanities has substantial potential to endow students with a deep understanding of the methods and techniques employed in this field, develop critical thinking and problem-solving skills, and promote active, project-based learning. Furthermore, it can offer a structured framework for the design of the curriculum itself, helping to ensure that it is comprehensive, logical, and aligned with learning outcomes.

Incorporating critical data literacy into digital humanities pedagogy represents a valuable opportunity to engage students in evaluating the implications of data modelling. Curricula can guide students to think critically about how selection, cleaning, and transformation of data embed certain assumptions and biases. Exercises in identifying exclusion or misrepresentation in dataset construction and model design help sensitize students to issues of epistemic injustice that may be obscured by technical processes (Irgens 2020). Further critical engagement involves assessing whose perspectives and narratives are privileged in data modelling pipelines, prompting reflection on power dynamics and marginalisation (D'Ignazio and Klein 2020). This ties into broader ethical questions around consent, access, and control over cultural data that students can unpack. Ultimately, cultivating skills to decode and interrogate the construction of data models, rather than passively accepting their authority, enables students to apply digital humanities tools thoughtfully.

CHALLENGES AND LIMITATIONS OF DATA MODELLING IN THE DIGITAL HUMANITIES

While data and knowledge modelling has been celebrated for its ability to structure, simplify and analyse complex systems in digital humanities, it is not without criticism. Scholars have pointed out potential issues ranging from the risk of oversimplification to the challenge of interpretation, and even ethical concerns around data usage and representation.

One key challenge is managing the complexity inherent in humanities data. This involves not only the complexity of the data itself but also the complexity of the socio-cultural phenomena it represents. This can be mitigated to an extent by adopting robust modelling techniques like semantic networks, ontologies, or machine learning models, which can handle high-dimensional, interconnected data. However, the challenge of adequately representing the intricacies of human experience within a model remains an open issue.

Managing complexity refers to the risk of oversimplification, which is one of the central concerns of digital humanities. When modelling complex phenomena, like social interactions or historical events, the process requires an element of abstraction, condensing reality into quantifiable parameters. For example, Johanna Drucker (2011) warns that this process might lead to overlooking nuances, complexities, and outliers, leading to misleading conclusions.

Ensuring data integrity is another concern. The quality of a model depends heavily on the quality of the input data. Unclear, missing, or incorrect data can distort the model's output. Techniques such as data cleaning, validation, and redundancy checks can help mitigate this issue, but perfect data integrity is rarely achievable, especially when dealing with historical or other hard-to-verify data sources.

Further critique emerges around the interpretation of models and their outputs. While models can help visualize data and reveal patterns, they cannot replace human judgment and critical thinking. David M. Berry articulates the necessity for a critical digital humanities, emphasising the potential for misinterpretation or manipulation of model results without proper understanding and contextual knowledge (Berry 2014). Techniques such as critical discourse analysis can provide systematic approaches to interpretation, but the inherent subjectivity of interpretation cannot be entirely eliminated. As Johanna Drucker argues, visualizations and models remain constructed representations, requiring contextualization (Drucker 2011). Data models cannot replace deep humanistic understanding. Similarly, generaliation in data models may fail to account for outliers and exceptions requiring close reading to identify (Piper 2018).

Beyond interpretational pitfalls, ethical issues arise in the realm of data usage and representation. In dealing with vast datasets, especially

those involving personal data, concerns around privacy, consent, and potential misuse of data come to the fore. Tara McPherson cautions against the racial and gender biases that can be present in coding and computational thinking, extending to data and knowledge modelling (McPherson 2012). Furthermore, the choice of which data to include or exclude in a model can reflect certain biases, which might lead to misrepresentation or marginalisation of certain groups or perspectives.

Critiques extend to the digital humanities' emphasis on quantitative methods, perceived by some as favouring scientific methods and the so-called "hard" sciences over traditional humanities disciplines, which often prioritize qualitative analysis and interpretive approaches. This concern is vocalized by Alan Liu, who contends that the critical cultural interpretation essential to the humanities should not be overshadowed by computational methods (Liu 2012).

While data modelling facilitates new research capabilities, it also warrants careful examination given its potential limitations. Scholars have argued computational techniques carry inherent biases that can propagate through data modelling pipelines (Benthall, Haynes 2019). Models trained on biased datasets may amplify distortive assumptions. Even in the absence of explicit biases, researchers caution that data models risk perpetrating "epistemic injustice" by flattening complex humanistic phenomena (Noble 2018). Generative text also risks perpetrating harmful biases embedded in training data. Models may reproduce stereotypical tropes or skew aggregate style representations towards overrepresented groups (Manela 2021). This necessitates critical assessment of model-generated content as constructed output requiring contextual interpretation. Ethical implications arise with text generation. Large language models trained on copyrighted data raise legal questions around creative ownership. Their ability to automate written content also warrants considering effects on human creativity and scholarship. Such models currently function as assistants, not autonomous creators or experts. Still, their interpretive limitations mean humanists must continue close reading, contextual analysis and cultivation of wisdom. By spurring critical discourse about AI's capacities and biases, large language models like GPT-4 demonstrate how digital humanities must guide emerging technologies towards ethical application in humanities enquiries. Other ethical concerns involve privacy, consent, and appropriate use of cultural data. Digital humanities research drawing on large datasets of personal information needs to implement safeguards around individual privacy and autonomy. As entities like archives and libraries digitise materials at scale, they must consider thorny issues around public access versus consent, as with indigenous community materials (Christen 2015).

Berry et al. (2019) argue that the digital humanities often takes a "signal processing" approach to cultural heritage, focusing on extract-

ing and manipulating data/signals while neglecting symbolic meaning and interpretation. This risks flattening complex phenomena into quantitative parameters and prioritising computation over humanistic interpretation. Data modelling enables new discoveries but should not lead to "signal processing" at the expense of humanistic questioning, ethics and symbolic understanding. Computational techniques require thoughtful application and decoding.

In summary, while there are methodologies and techniques that can help address some of the challenges associated with data and knowledge modelling in digital humanities, other issues, particularly those related to representation, bias, and interpretation, remain difficult to resolve entirely. These challenges highlight the importance of critical, reflective, and ethical practices in digital humanities. Ultimately, while enabling computational analysis, data modelling in the digital humanities demands a critical lens attending to biases, subjectivity, generalization, ethics, and context. Hybrid teams of humanists and computer scientists can thoughtfully apply modelling while avoiding pitfalls through transparent practices.

CONCLUSION

The methodological foundation of data, information, and knowledge modelling is integral to the field of digital humanities. It has not only enabled scholars to apply computational tools to humanities research but also opened up new avenues for discovery and understanding. These models, by providing a structured way to represent and analyse humanities data, have made research more systematic, scalable, and insightful.

The historical development, contemporary trends, and applications of data and knowledge modelling in various humanities disciplines attest to their transformative impact on research, analysis, and preservation of cultural heritage. They have not only served as crucial methodological foundations but also unlocked new dimensions of insight into the human experience and cultural heritage.

Looking to the future, advancements in artificial intelligence and machine learning promise to further enrich the complexity and depth of these models. As these technologies continue to evolve, data and knowledge modelling will undoubtedly play an increasingly vital role in shaping the future of digital humanities, driving new insights and understanding in the field.

This examination of data and knowledge modelling elucidates its vital yet complex role within digital humanities methodology. Modelling techniques enable computational analysis of cultural artifacts, dynamics and relationships at an unprecedented scale. By extracting patterns from corpora spanning genres, eras and languages, data

models reveal phenomena otherwise invisible to human scrutiny. However, as constructive simplifications, they risk flattening nuance and exception. Interpreting model outputs necessitates humanistic wisdom and scepticism towards generalisation.

Furthermore, employing data modelling in ethical, socially conscious ways remains imperative given dangers of perpetrating bias and other harms. As the digital humanities continue adopting advanced techniques like large language models, critical interrogation must accompany technical innovation. Ongoing progress will create immense opportunities for scholarship, but humanists must steward these tools towards expansive, equitable ends.

Ultimately, data and knowledge modelling establishes a framework to activate cultural data computationally while upholding principles of critical enquiry fundamental to the humanities. It enables asking new questions and marshalling new evidence. But humanistic study also connotes questioning how models construct reality, probing their constraints and biases. By maintaining this spirit of reflective application, the digital humanities can leverage data modelling for positive transformation while remaining grounded in humanistic ethics and understanding.

REFERENCES

Anthropic. 2023. "Introducing Claude." Anthropic blog. https://www.anthropic.com/index/introducing-claude

Benthall, Sebastian, and Bruce D. Haynes. 2019. "Racial Categories in Machine Learning." In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 289–98. New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287575

Berry, David M. 2014. *Critical Theory and the Digital*. London: Bloomsbury.

Berry, David M., M. Beatrice Fazi, Ben Roberts, and Alban Webb. 2019. "No Signal without Symbol: Decoding the Digital Humanities." In *Debates in the Digital Humanities 2019*, edited by Matthew K. Gold and Lauren F. Klein, 61–74. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/j.ctvg251hk.8

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. "On the Opportunities and Risks of Foundation Models." ArXiv 2108.07258. http://arxiv.org/abs/2108.07258

Borgman, Christine L. 2009. "The Digital Future Is Now: A Call to Action for the Humanities." *Digital Humanities Quarterly* 3 (4). http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. 2012. *Digital_Humanities*. Cambridge, MA: MIT Press.

Busa, Roberto. 1980. "The Annals of Humanities Computing: The Index Thomisticus." *Computers and the Humanities* 14 (2): 83–90. https://doi.org/10.1007/BF02403798

Christen, Kimberly. 2015. "Tribal Archives, Traditional Knowledge, and Local Contexts: Why the 's' Matters." *Journal of Western Archives* 6 (1). https://doi.org/10.26077/78d5-47cf

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle. 2018. "Modelling: Thinking in Practice. An Introduction." *Historical Social Research Supplement* 31: 7–29. https://doi.org/10.12759/hsr.suppl.31.2018.7-29

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." ArXiv 1810.04805. http://arxiv.org/abs/1810.04805

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Cambridge: MIT Press.

Doorn, Peter. 2021. "Historical Databases, Big and Small." *Historical Life Course Studies* 10: 24–29. https://doi.org/10.51964/hlcs9562

Drucker, Johanna. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5 (1). http://digitalhumanities.org:8081/dhq/vol/5/1/000091/000091.html

Drucker, Johanna. 2014. Graphesis: Visual Forms of Knowledge Production. Cambridge: Harvard University Press.

Drucker, Johanna. 2021. *The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Scholarship*. Abingdon: Routledge.

Edmond, Jennifer. 2018. "Managing Uncertainty in the Humanities: Digital and Analogue Approaches." In *TEEM18: Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 840–844. https://doi.org/10.1145/3284179.3284326

Elgammal, Ahmed, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. "CAN: Creative Adversarial Networks, Generating 'Art' by Learning About Styles and Deviating from Style Norms." *ArXiv* 1706.07068. http://arxiv.org/abs/1706.07068

Fiormonte, Domenico, Teresa Numerico, and Francesca Tomasi. 2015. *The Digital Humanist*. Translated by Desmond Schmidt and Christopher Ferguson. Brooklyn: Punctum.

Flanders, Julia, and Fotis Jannidis. 2015. "Knowledge Organization and Data Modeling in the Humanities." White Paper. https://www.wwp.northeastern.edu/outreach/conference/kodm2012/flanders_jannidis_datamodeling.pdf.

Fogel, Robert William, and Stanley L. Engerman. 1974. *Time on the Cross: The Economics of American Negro Slavery*. Boston: Little and Brown.

Gagarina, Dinara, and Sergei Kornienko. 2018. "Introduction to Digital Humanities: Course Based on Data Types." In *Proceedings of the First EADH Conference: "Data in Digital Humanities"*. Galway: NUI Galway. https://eadh2018.exordo.com/files/papers/151/final_draft/gagarina-eadh-final.pdf

Gold, Matthew K., ed. 2012. Debates in the Digital Humanities. Minneapolis: University of Minnesota Press.

Gold, Matthew K., and Lauren F. Klein, eds. 2019. *Debates in the Digital Humanities 2019*. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/j.ctvg251hk

Ide, Nancy, and David Woolner. 2007. "Historical Ontologies." In *Words and Intelligence II*, edited by Thierry Poibeau, Horacio Saggion, Jakub Piskorski, and Roman Yangarber, 63–79. Dordrecht: Springer Netherlands. https://doi.org/10.1007/1-4020-5833-0_7

Irgens, Golnaz Arastoopour, Simon Knight, Alyssa Friend Wise, Thomas M. Philip, Maria C. Olivares, Sarah Van Wart, Sepehr Vakil, Jessica Marshall, Tapan Parikh, M. Lisette Lopez, et al. 2020. "Data Literacies and Social Justice: Exploring Critical Data Literacies Through Sociocultural Perspectives." In *Proceedings of the 2020 International Conference of the Learning Sciences*. Nashville: International Society of the Learning Sciences (ISLS).

Isaksen, Leif, Rainer Simon, Elton T.E. Barker, and Pau de Soto Cañamares. 2014. "Pelagios and the Emerging Graph of Ancient World Data." In *Proceedings of the 2014 ACM Conference on Web Science*, 197–201. New York: Association for Computing Machinery. https://doi.org/10.1145/2615569.2615693

Jannidis, Fotis. 2018. "Modeling in the Digital Humanities: A Research Program?" *Historical Social Research Supplement* 31: 96–100. https://doi.org/10.12759/hsr.suppl.31.2018.96-100

Jockers, Matthew L. 2013. Macroanalysis: Digital Methods and Literary History. Topics in the Digital Humanities. Urbana: University of Illinois Press.

Kim, Evgeny, Sebastian Padó, and Roman Klinger. 2017. "Investigating the Relationship between Literary Genres and Emotional Plot Development." In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 17–26. Vancouver: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2203

Kitchin, Rob, and Gavin McArdle. 2016. "What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets." *Big Data & Society* 3 (1). https://doi.org/10.1177/2053951716631130

Knuth, Donald Ervin. 1968. *The Art of Computer Programming*. Vol. 1. Reading: Addison-Wesley.

Liu, Alan. 2012. "The State of the Digital Humanities: A Report and a Critique." *Arts and Humanities in Higher Education* 11 (1–2): 8–41. https://doi.org/10.1177/1474022211427364

Luccioni, Alexandre, and Anna Rogers. 2023. "Mind Your Language (Model): Fact-Checking LLMs and Their Role in NLP Research and Practice." *ArXiv* 2308.07120. http://arxiv.org/abs/2308.07120.

Manela, Daniel de Vassimon, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. "Stereotype and Skew: Quantifying Gender Bias in Pre-Trained and Fine-Tuned Language Models." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2232–42. https://aclanthology.org/2021.eacl-main.190/.

Manovich, Lev. 2013. *Software Takes Command: Extending the Language of New Media*. New York: Bloomsbury.

Manovich, Lev. 2020. *Cultural Analytics*. Cambridge, MA: MIT Press.

McCarty, Willard. 2004. "Modeling: A Study in Words and Meanings." In A Companion to Digital Humanities, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell. http://www.digitalhumanities.org/companion/

McCarty, Willard. 2005. *Humanities Computing*. Basingstoke: Palgrave Macmillan.

McCarty, Willard. 2018. "Modelling What There Is: Ontologising in a Multidimensional World." *Historical Social Research Supplement* 31: 33–45. https://doi.org/10.12759/hsr.suppl.31.2018.33-45

McPherson, Tara. 2012. "Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 139–60. Minneapolis: University of Minnesota Press.

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank van Harmelen. 2014. "Semantic Technologies for Historical Research: A Survey." *Semantic Web* 6 (6): 539–64. https://doi.org/10.3233/SW-140158

Minsky, Marvin Lee. 1986. *The Society of Mind*. New York: Simon and Schuster.

Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Mylonas, Elli, Gregory Crane, Kenneth Morrell, and D. Neel Smith. 1993. "The Perseus Project: Data in the electronic age." In *Accessing Antiquity: The Computerization of Classical Studies*, 133–157. Tucson: University of Arizona Press.

Noble, Safiya Umoja. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. New York: New York University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Piper, Andrew. 2018. *Enumerations: Data and Literary Study*. Chicago: The University of Chicago Press.

Riley, Jenn. 2017. *Understanding Metadata: What Is Metadata, and What Is It For? A Primer.* Baltimore: NISO Press.

Rockwell, Geoffrey, and Stéfan Sinclair. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, MA: MIT Press.

Romanello, Matteo, and Michele Pasin. 2013. "Citations and Annotations in Classics: Old Problems and New Perspectives." In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*, 1–8. New York: Association for Computing Machinery. https://dl.acm.org/doi/10.1145/2517978.2517981

Rosenbloom, Paul S. 2012. *On Computing: The Fourth Great Scientific Domain*. Cambridge, MA: MIT Press.

Shannon, Claude Elwood. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sperberg-McQueen, C. M., and Lou Burnard, eds. 1994. *Guidelines for Electronic Text Encoding and Interchange*. Chicago: Text Encoding Initiative. https://tei-c.org/Vault/GL/p4beta.pdf

Spiro, Lisa. 2012. "'This Is Why We Fight': Defining the Values of the Digital Humanities." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 16–35. Minneapolis: University of Minnesota Press.

Svensson, Patrik. 2010. "The Landscape of Digital Humanities." *Digital Humanities Quarterly* 4 (1). http://www.digitalhumanities.org/dhq/vol/4/1/000080/000080.html

Thaller, Manfred. 1985. "Beyond Collecting – On the Design and Implementation of CLIO, a DBMS for the Historical Sciences." In *Data Bases in the Humanities and the Social Sciences* 2, edited by Robert F. Allen, 443–68. Osprey: IDEA Publications.

Thaller, Manfred. 2012. "Controversies around the Digital Humanities: An Agenda." *Historical Social Research / Historische Sozialforschung* 37 (3 (141)): 7–23. https://www.jstor.org/stable/i40077811.

Turing, A. M. 1937. "On Computable Numbers, with an Application to the Entscheidungsproblem." In *Proceedings of the London Mathematical Society*, s2-42 (1): 230–65. https://doi.org/10.1112/plms/s2-42.1.230

Underwood, Ted. 2014. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127 (1): 64–72. https://doi.org/10.1525/rep.2014.127.1.64.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago The University of Chicago Press.

Unsworth, John. 2000. "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" Presented at the Symposium on Humanities Computing: Formal Methods, Experimental Practice, King's College London, May 13. https://people.brandeis.edu/ũnsworth/Kings.5-00/primitives.html.

von Neumann, John. 1945. "First Draft of a Report on the EDVAC." Contract No. W-670-ORD-4926 between the United States Army Ordnance Department and the University of Pennsylvania. https://web.mit.edu/STS.035/www/PDFs/edvac.pdf.

Weingart, Scott B. 2011. "Demystifying Networks, Parts I & II." *Journal of Digital Humanities* 1 (1). http://journalofdigitalhumanities.org/1-1/demystifying-networks-by-scott-weingart/.