DHQ: Digital Humanities Quarterly
Preview
2024
Volume 18 Number 3

Manuscript Catalogues as Data for Research: From Provenance to Data Decolonisation

Huw Jones <hej23_at_cam_dot_ac_dot_uk>, Cambridge University Library; Cambridge Digital Humanities https://orcid.org/0000-0002-8533-9083

Yasmin Faghihi <yf227_at_cam_dot_ac_dot_uk>, Cambridge University Library; Cambridge Digital Humanities https://orcid.org/0000-0001-5556-168X

Abstract

This paper discusses a recent project which applied computational methods to catalogue data in an attempt to generate new information on the provenance of Islamicate manuscripts in UK repositories. Using a subset of records taken from the Fihrist Union Catalogue as a dataset, we analysed and grouped together manuscript descriptions which shared selected physical features, then examined the occurrence of records with secure provenance within those groups to see if information on the place of origin could be extrapolated across them. While we gained useful information regarding the provenance of manuscripts, the chief conclusion of the project was that catalogue data, in its current state, poses serious challenges for quantitative analysis. This is partly due to the various purposes for which data has traditionally been collected (in contexts where codicological descriptions had a different purpose) and partly due to inconsistencies in the dataset. In our conclusion we put forward strategies for working with inconsistent data, make suggestions for changes to cataloguing practices to answer the requirements of digital methods, and propose new research questions addressing the history of catalogues and cataloguing practices which came into focus during the project. We also make a case for the potential of digital methods to enable new approaches to decolonisation, focusing on data modelling, data provenance, and accessibility.

1 Introduction

In *Islamic Codicology*, François Déroche calls upon the creators of manuscript catalogues to produce “... coherent sets of documents that shed light on one another” [Déroche 2006, 17]. Given the complex relationships between historical manuscript cultures and collections of manuscripts, between collections of manuscripts and the catalogues which describe them, and between catalogues as datasets and the outputs of digital methods, what is the nature of this “light”? What new insights do we gain by using data derived from manuscript catalogues to generate new information at scale? And in doing so, what do we learn about historical manuscript cultures, about the history and formation of collections of manuscripts, about the processes by which catalogues are created, and about our own motivations as expressed in our digital methodologies? 1

This paper discusses a recent project, funded by the Cambridge Humanities Research Grants scheme, which applied computational methods to catalogue data in an attempt to “shed light” on the provenance of Islamicate manuscripts in UK repositories. While its focus was on manuscript history, the project was created in the context of contemporary concerns around the decolonisation of collections in cultural heritage institutions, which have given a new importance to research on the origins of manuscripts in UK collections. However, establishing the provenance of Islamicate manuscripts presents particular challenges for both cultural and historical reasons. The places and dates of production are rarely mentioned in the texts themselves — and tracing the biography of Islamicate manuscripts held in UK repositories is further complicated not only by the texts' migration to the UK but also their previous movement within the 2

Islamicate world. As a result, we rely heavily on the physical features of manuscripts when attempting to ascertain their place and date of production as well as their wider history.

Using a subset of records taken from the Fihrist Union Catalogue for Manuscripts from the Islamicate World in UK Repositories as a dataset, we used computational methods to analyse and group together manuscript descriptions which shared selected physical features (size, script, orientation etc.), then examined the occurrence of records with secure provenance within these groups to see if information on the place of origin could be extrapolated across them. We also investigated how new information generated by this process could be integrated back into the dataset, while retaining a clear sense of its own data provenance. Although we gained useful information regarding the provenance of manuscripts, the chief conclusion of the project was that catalogue data, in its current state, poses serious challenges for quantitative methods. However, our engagement in the process also raised interesting questions about the history and nature of the collections described, about cataloguing practices and data modelling, and about the relationship between digital methods and research questions. 3

In this paper we present our activities as a case study for digital approaches to manuscript catalogues — from collections through catalogues and datasets, to computational methods and their results. In our conclusion we put forward strategies for working with inconsistent data, make suggestions for changes to cataloguing practices to answer to the requirements of digital methods, and propose new research questions on the history of catalogues and cataloguing practices which came into focus during the project. We also make a case for the potential of digital methods to enable new approaches to decolonisation, focusing on data modelling, provenance, and accessibility. 4

2 Codicology and Islamicate Manuscripts

There are a number of reasons for the existence of large collections of Islamicate manuscripts in UK repositories. First, the Islamicate manuscript tradition is in itself very large, partly due to the early adoption of paper as writing support (from the 8th century onwards) and partly due to the late adoption of printing at industrial scale, which led to the survival of a tradition of copying manuscripts by hand into the 20th century. This is further amplified by the historical expansion of Islam and the subsequent dominance of Arabic as the language of scholarship across a large geographical area. We therefore find a vast number of handwritten texts, some of which have travelled on trade or pilgrimage routes, while others remained in their area of production. In a UK context, empire, trade, and missionary activities led to a long-standing engagement with Islam and the Islamicate world, which in turn led to the development of large manuscript collections. 5

Collections in Western libraries and archives, which are the focus of this study, are typically made up of items from across the Islamic world. Here we find related texts, copied in different countries, which bear typical features of the craft, materials, and artwork relating to the local environment in which they were produced (the Qur'an being perhaps the best example of a text which was produced extensively across a wide geographical area). While the text has traditionally been the basis of philological and bibliographic research, a focus on the physicality of the book as a carrier of text allows us to investigate the material culture both of the texts themselves (as different texts tend to have different material contexts), and of the contexts which produced them. 6

With the rise of Islamic codicology as an evolving discipline, physical aspects of manuscripts are attracting more attention, especially in a Western context. Western collections of Islamicate manuscripts often contain a diverse mixture of traditions both in terms of places of origin and eras of production. Compared to bodies of material which have remained in their original contexts, these collections can appear idiosyncratic and require a different approach when investigating their origins and subsequent histories. Given the comparative lack of indications of geographical origins in the text of Islamicate manuscripts, analysis of physical features has become a key method for establishing provenance. 7

Manuscript collections from the Islamicate world demonstrate an enormous variety of physical features in terms of shapes, orientations, materials, layouts, and binding. The long history of paper production and use in the Middle East adds an additional layer of complexity to the texts' codicological makeup. While in their current context the places of origin have often been obscured, these manuscripts have nevertheless been produced within a coherent tradition. Codicology gives us a framework to establish criteria for comparative observation and analysis which can give clues to 8 these traditions. This framework, captured in machine-readable form, provides the basis for a flexible and expandable data model which can respond to new discoveries and approaches.

3 Quantitative Codicology and Manuscript Catalogues

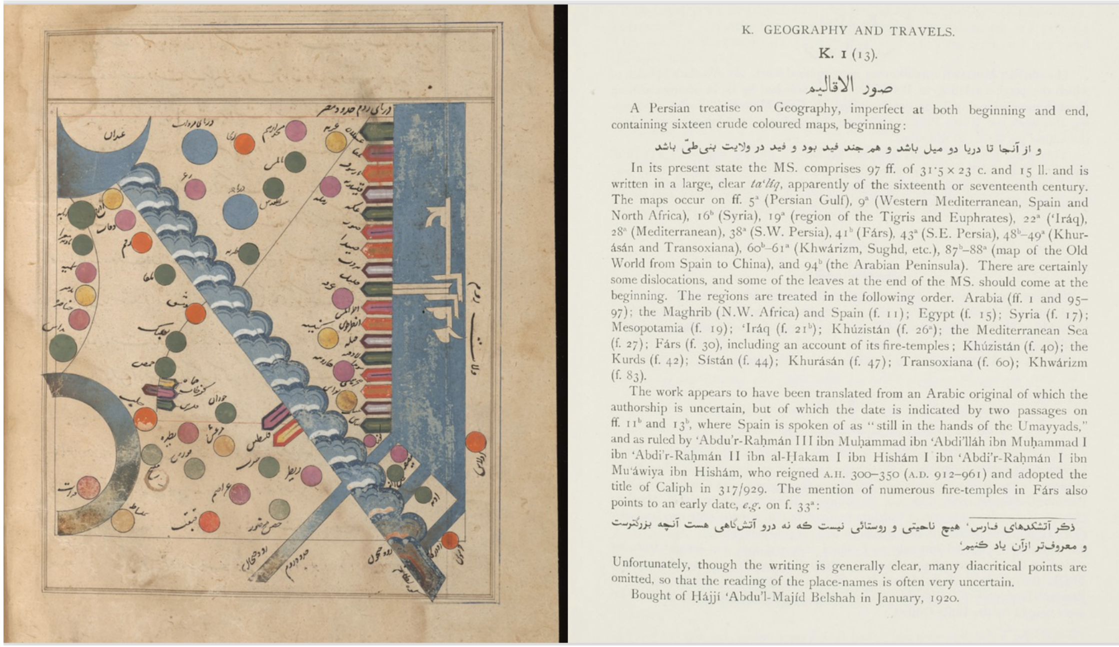
Quantitative codicology is concerned chiefly with the statistical analysis of the correlations between a subset of physical characteristics of a corpus, as well as what the relationships between those characteristics might tell us about a) the items themselves, and b) the social, economic, and cultural contexts of their production [Ornato 2020]. In addition to studies based on particular manuscript cultures, quantitative codicology has been widely applied in comparative research that attempts to isolate features that are the particular result of the context of production (from those which are common to all contexts) and to trace the effect of cross-cultural influences and borrowings on manuscript cultures [Beit-Arié 2020]. Existing computational work has tended to follow this model, concentrating either on what can be inferred from relationships within a single dataset, or on combining information from two or more datasets to generate new information. 9

One of the chief challenges of quantitative work on manuscript collections is the tension between individualising and generalising modes of enquiry raised by Daston and Galison in the context of the history of science [Daston and Galison 2007]. Although our methodology is largely based on shared features of manuscripts, manuscripts themselves are, by their very nature, unique objects which resist general approaches. Burrows has pointed to the dominance of individual or small-scale approaches in manuscript studies [Burrows 2018], but, as Daston and Galison point out, there are features which only reveal themselves at scale: 10

> Some significant historical phenomena are invisible at the local level, even if their manifestations must by definition be located somewhere, sometime. There are developments that unfold on a temporal and geographic scale that can only be recognised at a local level once they have been spotted from a more global perspective. Just as no localized observer alone can detect the shape of a storm front or the distribution of an organic species, so some historical phenomena can be discerned only by integrating information from a spread of contexts. [Daston and Galison 2007, 47]

Where can we find codicological information at sufficient scale to begin to detect the shape and internal relationships of our collections? Scientific analysis of physical aspects of manuscripts, such as watermark imaging and analysis, XRTF, multi-spectroscopy, and radiocarbon dating, produce interesting results for individual manuscripts but are expensive and difficult to apply at scale. Automated analysis of images, facilitated by IIIF, has had some success in identifying physical features such as script type and number of lines per page (see for example [Chandna et al. 2016], [Van Lit 2019]), but even in the age of mass digitisation the proportion of the corpus which has been imaged remains comparatively small and unevenly distributed [Zaagsma 2022]. Even for more easily accessible characteristics such as script, material, size, illustration, and binding, the effort involved in assembling a new dataset at any scale through inspection of manuscripts can be proscriptive, especially for dispersed collections. As Marilena Maniaci notes: “Given the practical impossibility of directly examining many hundreds of manuscripts, one has to accept the necessity of substituting the task with a census of data collected from previously published descriptions” [Maniaci 2022, 467]. 11

This leads us to existing manuscript catalogues as a source of data for quantitative approaches, but manuscript catalogues are far from unproblematic as datasets, “compris[ing] subsets of records written at different times, by different people, according to different principles and goals” [Baker, Salway, and Roman 2022, 3]. Or, in Camillo Formigatti's neat formulation, “[c]atalogues give manuscripts a voice, but the language in which they speak varies according to the interests and priorities of the scholars who catalogue them” [Formigatti 2017, 29]. In a union catalogue such as Fihrist, drawn together from many sources created in a variety of contexts over an extended period of time, these voices multiply. 12



K. GEOGRAPHY AND TRAVELS.

K. 1 (13).

صور الاقاليم

A Persian treatise on Geography, imperfect at both beginning and end, containing sixteen crude coloured maps, beginning:

و از آنجا تا دریا دو میل باشد و هم جند فید بود و فید در ولایت بنی‌طیّ باشد

In its present state the MS. comprises 97 ff. of 31·5 × 23 c. and 15 ll. and is written in a large, clear *ta'líq*, apparently of the sixteenth or seventeenth century. The maps occur on ff. 5^{a} (Persian Gulf), 9^{a} (Western Mediterranean, Spain and North Africa), 16^{b} (Syria), 19^{a} (region of the Tigris and Euphrates), 22^{a} ('Iráq), 28^{a} (Mediterranean), 38^{a} (S.W. Persia), 41^{b} (Fárs), 43^{a} (S.E. Persia), 48^{b}–49^{a} (Khurásán and Transoxiana), 60^{b}–61^{a} (Khwárizm, Sughd, etc.), 87^{b}–88^{a} (map of the Old World from Spain to China), and 94^{b} (the Arabian Peninsula). There are certainly some dislocations, and some of the leaves at the end of the MS. should come at the beginning. The regions are treated in the following order. Arabia (ff. 1 and 95–97); the Maghrib (N.W. Africa) and Spain (f. 11); Egypt (f. 15); Syria (f. 17); Mesopotamia (f. 19); 'Iráq (f. 21^{b}); Khúzistán (f. 26^{a}); the Mediterranean Sea (f. 27); Fárs (f. 30), including an account of its fire-temples; Khúzistán (f. 40); the Kurds (f. 42); Sístán (f. 44); Khurásán (f. 47); Transoxiana (f. 60); Khwárizm (f. 83).

The work appears to have been translated from an Arabic original of which the authorship is uncertain, but of which the date is indicated by two passages on ff. 11^{b} and 13^{b}, where Spain is spoken of as "still in the hands of the Umayyads," and as ruled by 'Abdu'r-Raḥmán III ibn Muḥammad ibn 'Abdi'lláh ibn Muḥammad I ibn 'Abdi'r-Raḥmán II ibn al-Ḥakam I ibn Hishám I ibn 'Abdi'r-Raḥmán I ibn Mu'áwiya ibn Hishám, who reigned A.H. 300–350 (A.D. 912–961) and adopted the title of Caliph in 317/929. The mention of numerous fire-temples in Fárs also points to an early date, *e.g.* on f. 33^{a}:

ذکر آتشکدهای فارس، هیچ ناحیتی و روستائی نیست که نه درو آتش‌گاهی هست آنچه بزرگترست
و معروف‌تر ازآن یاد کنیم،

Unfortunately, though the writing is generally clear, many diacritical points are omitted, so that the reading of the place-names is often very uncertain.

Bought of Ḥájjí 'Abdu'l-Majíd Belshah in January, 1920.

Figure 1. Digitised image and catalogue record for K.1, Cambridge University Library, recording point of acquisition but not earlier history. Note the use of language: "crude coloured maps".

Behind the catalogues, each with their different voice, lie collections — and it is debatable how far contemporary collections of manuscripts in Western repositories can be thought of as representative of historical manuscript cultures. In addition to local factors in the formation and survival of specific collections, such as the interests of collectors and institutions, there are issues of "completeness" at a more fundamental level: 13

> It is necessary to take into account that the extant manuscript patrimony is but a small fraction of the quantity of volumes produced; that the volumes having a date are only a small part of the extant patrimony; that neither the general cataloguing of the extant patrimony nor that of the dated manuscripts is exhaustive at the present time. This multistage filter created by the destructive work of time and by the inevitable gaps in reporting would not be significant if the reduction of the population at each step was uniform relative to the initial state. But this is not the case. [Ornato 2020, 656]

It is tempting to label the manuscript record "incomplete", but that would miss the point — it is complete on its own terms, but the nature of its completeness cannot be neatly packaged and understood in relation to the record's survival and loss. Instead, the record speaks to an interplay of multiple factors over time, each of which might merit our interest, which have "completed" the archive in its current state. The archive we have is the archive we have — or, as Carolyn Steedman would have it, "[i]n the Archive, you cannot be shocked at its exclusions, its emptiness, at what is not catalogued … its condition of being deflects outrage" [Steedman 1998, 68]. 14

Given the complex nature of both manuscript collections and the catalogues which describe them, what new information can we usefully derive from computational approaches to data derived from manuscript catalogues? We followed Marilena Maniaci's pragmatic advice to give up on any "ill-founded attempt at achieving 'objectiveness'" and instead accept "... the limits set by the adopted approach, the nature of surviving documentation, and the quantity and quality of the data upon which the analysis is carried out" as integral to our research [Maniaci 2022, 467]. In a sense, the most interesting thing about manuscript catalogue data is its very complexity — and the way in which it leads us back down through the whole long stack of cataloguers and institutions, collectors and collections, data and method, loss and acquisition. Concrete results from quantitative methods on manuscript collections do have value to researchers but need to be understood in the broader context of all the many processes and actors which have led to the formation of the evidence we use to produce them. 15

4 Related Research

4.1 Ontology-Based Analysis of the Large Collection of Historical Hebrew Manuscripts

Projects using manuscript catalogue data to generate new information have tended (with some success) to focus on the pragmatic business of filling in gaps in our knowledge rather than on the complex nature of the data and its relation to the “realities” of historical manuscript cultures. Zhitomirsky-Geffet, Prebor, and Miller extracted data from the National Library of Israel's catalogue system (“the largest catalogue in the world of Hebrew manuscripts” [Zhitomirsky-Geffet, Prebor, and Miller 2020, 688]), using an event-based ontology to “complete missing data by inference from the constructed ontology and from external resources” and to perform a quantitative analysis of the catalogue as a whole [Zhitomirsky-Geffet, Prebor, and Miller 2020, 689]. The new data produced by the project was generated a) by reference to external sources,[1] and b) by inference within the dataset.[2]. 16

The project differed from ours its concentration on texts and people (scribes and authors) rather than physical characteristics of manuscripts and in its reliance on direct inferences and references (internal and external to the dataset) based on existing information about single, known features of manuscripts rather than on groups of features in combination. While some attention was paid to the sources and nature of the catalogue data which underpinned the project, this mainly concerned “inconsistencies” and “errors” in the data which are “... explained by the fact that the catalogue has been created manually by dozens of different cataloguers during the period of over 70 years” [Zhitomirsky-Geffet, Prebor, and Miller 2020, 689]. The different attitudes, interests, and biases that may have informed cataloguing practices throughout this period are largely glossed over. 17

The research was framed as an investigation of the manuscript collection itself, rather than an attempt to make large claims about its actual relation to historical manuscript culture, so a detailed discussion of the historical formation of the collection was possibly out of scope. The comprehensive and impressive section which deals with quantitative analysis of the relationships between certain aspects of the data (for instance, the “most copied authors in different time periods and countries”) does point to interesting potential research avenues where the completeness and nature of the surviving manuscript record as manifested in the collection would take on more importance. The influence of “the accidental survival of Hebrew manuscripts over time” [Zhitomirsky-Geffet, Prebor, and Miller 2020, 716] is acknowledged in their diachronic analysis of script types in the same dataset, which is directly relevant to our own larger research ambitions in highlighting the importance of some manuscript features, not only in localising the place of production, but also in tracing the movement of manuscript cultures over time — and in looking at how these movements might relate to contemporary political, social, environmental, and economic factors [Zhitomirsky-Geffet, Prebor, and Miller 2020]. However, we would argue that few of the factors that lead to the survival of manuscripts in collections are “accidental”. 18

4.2 *Mapping Manuscript Migrations*

The *Mapping Manuscript Migrations* project took the history of collections and collectors as its primary focus, combining four large datasets as the basis of an investigation into the provenance of manuscripts — that is “who owned them, in which places, and at which times” *after* the moment of their initial production: 19

> Each object has usually been part of a series of collections over its lifetime, and this movement of objects between collections has its own history. Similarly, each collection has its own history of formation and (usually) dispersal, depending on whether the collectors were individuals, private institutions, or modern public institutions … The network of relationships between people and institutions involved in the ownership and transmission of cultural collections can also reveal a good deal about the more general networks of influence and social and political relationships in a particular society. [Burrows 2018, 5]

While the project had its own specific research questions relating to the movement of manuscripts over time, it was also preoccupied with general issues regarding the analysis of manuscript catalogue data at scale — in particular “a lack of coherent, interoperable infrastructure for the data relating to these manuscripts” — which makes the visualisation and 20 analysis of the history and movement of manuscripts "… difficult — if not impossible — to carry out" [Burrows 2018, 2]. In this context and building on previous work relating to the manuscript collection of a single collector, Sir Thomas Phillips, the focus was on the modelling and construction of a new linked open data resource derived from four datasets: the *Schoenberg Database of Manuscripts*, the *Medieval Manuscripts in Oxford Libraries* catalogue, and the *Bibale* and *Medium* databases of the Institut de Recherche et d'Histoire des Textes.

As each of these datasets used a different data standard, crosswalking between formats was a core part of the project, with one particular mapping, from TEI into RDF, described in detail in a separate paper [Burrows et al. 2021]. However, the work also highlighted fundamental problems in establishing identity, not just of the manuscripts themselves, but also for entities described in manuscript catalogues. Burrows focuses on people and works, but our own research encountered similar difficulties in finding robust identification schemes or consistent vocabularies for describing physical aspects of manuscripts such as layout, script, material, and binding. What we learned from this project is that variations in metadata format are less problematic in the analysis of manuscript catalogue data than variations in the identification and description of specific aspects of manuscripts — a factor which came into sharp focus in our own work. 21

4.3 *Legacies of Catalogue Descriptions and Curatorial Voice*

Mapping Manuscript Migrations plays close attention to the history of collections but does not provide much discussion on the motivations or context of those who have described them, or how these are reflected in the manuscript catalogues we use as the basis for our research. This was the subject of a recent project, *Legacies of Catalogue Descriptions and Curatorial Voice*, funded under the Towards a National Collection programme. *Legacies* highlighted the specific circumstances under which catalogues are created, edited, and reused — particularly in the context of the reuse of catalogue data at scale as datasets in the digital humanities: "Crucially, the historically specific labours and practices of catalogue production are all too easily obscured by the presentation of the catalogue as an always-already present unifying entity, and are further obscured when collections are federated for access, into datasets, or as machine readable endpoints. [Baker, Salway, and Roman 2022, 2]" 22

The project looked at the history of the transmission of the "voice" of a single cataloguer, Mary Dorothy George, over the course of almost a century — from "printed volumes to networked digital data". Using automated methods, they were able to trace the transmission of this voice across a series of catalogues and attempt to piece together the context of the original descriptions and of the multiple acts of revision and adaptation to which they had been subject. This work has specific relevance to our dataset, which is an aggregation of legacy data and new cataloguing generated by multiple institutions over a long period. The project emphasises the additional importance of data provenance when datasets are used as the basis of computational analysis, or indeed as training data for neural networks. In our conclusion we discuss research avenues where variations in cataloguing practices (and intentions) are not only a factor in the resolution of disparate datasets but also an object of study in their own right. 23

4.4 *Bibliotheca Arabica*

Also of relevance to our project, both in terms of subject matter and approach, is the *Bibliotheca Arabica* [Kinitz and Efer 2023], a long-term project with a focus on manuscript notes and their value in discerning the circulation and origin of the manuscript in its social and historical context. Manuscript notes are important codicological features found in many manuscripts within the Islamicate and related traditions which tell us much about the history of manuscripts after the point of production. Using graph database technologies, this project identifies people and collections through their relationship to dates, works, and objects. The ambitious scope of this project, including data from collections at scale and from a variety of sources, could provide further evidence in corroboration with data derived from physical descriptions. We have already had preliminary conversations on data standardisation between our projects. 24

5 Dataset

5.1 Fihrist

Launched in 2011, Fihrist originated as a JISC-funded collaborative project between Cambridge University Library and 25 the Bodleian Libraries, Oxford to build an online catalogue for manuscripts from the Islamicate world. Descriptions and indexes of manuscripts in these collections had previously only existed in out-of-print catalogues, handlists, and card indexes, some dating back to the 18th century. The idiosyncratic structuring of information and lack of accessibility in these resources made them difficult to use and navigate. The aim of the project was to find an open, digital solution to the issue of discovering both the manuscripts themselves and the texts they contained, which went beyond the limitations of bibliographic cataloguing as applied to modern publications.

An additional goal was to create a functional and sustainable infrastructure which could bring together collections held in different repositories, thus facilitating access and discovery across large datasets. To achieve this, we embarked on the process of manually converting descriptions from existing catalogues into standardised data and organising the resulting records in a flexible and expandable data model which could incorporate further information over time. An interface with selected search and browse functions was developed for discovery, and the data set was made openly accessible on GitHub. 26

Following further funding from JISC, Cambridge Digital Humanities, and the Cambridge Humanities Research Grants scheme, Fihrist has grown to be a union catalogue for manuscripts from the Islamicate world in UK repositories, containing over 15,000 manuscript descriptions from 22 institutions. While many of these records are based on legacy data drawn from existing catalogues, the data in Fihrist has also been enriched by descriptions resulting from new research and cataloguing projects. The most obvious point of interaction with the resource is through the web interface, which has done much to facilitate discovery of and access to the manuscript material. However, in this article, our focus is on Fihrist as a dataset for research — and how using the catalogue records as data led us to critically examine the type of information traditionally collected in manuscript catalogues, as well as the correlation between selection, structure, description practices, and an evolving research landscape. 27

A crucial factor in Fihrist's development was the choice of TEI as a data standard. TEI is, according to its own webpage, “a consortium which collectively develops and maintains a standard for the representation of texts in digital form”, expressed in a set of guidelines for the encoding and description of text. The fact that Fihrist was initially developed as a stand-alone resource meant that it did not have to conform to the restrictions imposed by existing library systems, which tended to use more restrictive metadata standards such as MARC or EAD. This allowed us to take advantage of the flexibility and expandability of TEI in integrating manuscript research into the descriptive records. 28

Equally important was the adoption of some of the established standards, classifications, and taxonomies used by libraries and archives. While providing a set of lightweight hierarchical structures, TEI is broadly agnostic about the contents of elements and attributes. Reusing bibliographic standards for names and subjects, abbreviations for languages, and standardised transliteration systems was intended to facilitate the creation of content that was both internally consistent and compatible with that found in other datasets — in particular Anglo-American-language-based libraries and archives. Fihrist also maintains its own internal authority files for names, works, and subjects, which are linked where possible to external identifier schemes such as the VIAF and the Library of Congress, enabling future linked open data approaches. 29

5.2 TEI

With its ability to capture complex data for both publication and analysis, “[t]he Manuscript Description section of TEI Guidelines has become the de-facto schema for structuring detailed descriptions of manuscripts” [Burrows et al. 2021, 1]. For catalogue data it crucially allows us to capture narrative information in combination with encoded values suitable for quantitative approaches. While TEI has been criticised for offering “various ways of encoding the same basic information, with no definitive agreed standards” [Burrows et al. 2021, 1], in this respect it simply reflects a fundamental issue in data-driven digital humanities: 30

> Some of the most fertile and urgent areas of digital humanities research involve the question of how to develop data modelling approaches that accommodate both the self-reflexivity required by humanities research and the actionability and computational clarity required by the digital domain.
> [Flanders and Jannidis 2015, 236]

While TEI is not without its drawbacks, many of the criticisms aimed at it are misplaced (see e.g., [Cummings 2019]). The complexities around variation in practice, intention, and vocabulary which became clear in our project are real issues arising from manuscript cataloguing itself, not necessarily inherent in TEI. The open and flexible nature of TEI simply provides a framework for these issues to be discussed, clarified, and recorded. A number of scholars (including ourselves) have written on TEI's role as a research community rather than a metadata standard (e.g., [Mylonas and Renear 1999], [Cummings 2019], [Faghihi, Holford, and Jones 2022]), and its guidelines have evolved and continue to evolve in response to new materials and methodologies. 31

Our own collaborative activities have been centred on the development of a consolidated schema for manuscript description that forms the data model of a number of digital libraries, manuscript catalogues, and large, interoperable datasets for research. 32

When we first started to use TEI for manuscript description in the OCIMCO project to create a combined data set for Islamicate manuscripts from Cambridge and Oxford, we had no prototype other than the schema generated for Enrich, a project focused on the description and encoding of European manuscript resources. While this served as a basis, it was felt that a more customised approach was required for Islamicate manuscripts, following an established mindset focused on the differences between manuscript cultures rather than what they might have in common. A schema was commissioned which contained specialised elements for “distinct” features, such as the encoding of patronymic names and ruling practices. 33

However, while working with this schema, we came to the conclusion that we had made a conceptual error and that a standardised and consolidated approach across manuscript cultures was a more effective data model which would foster the kind of interdisciplinary work we were interested in. This approach focussed on the remodelling of data to make it responsive to digital methods and research questions, rather than trying to replicate the organisation and structures present in the legacy data. Our consolidated approach has generated interesting discussions about the feasibility and/or desirability of using a single schema a) across manuscript cultures, and b) to answer a variety of research questions. We address some of these issues in our conclusion. 34

One fundamental feature of our dataset is worth exploring in more detail. TEI allows you to use the Manuscript Description (msDesc) module either in the header (the descriptive metadata part of the record) or in the `<text>` element (as a transcription of the source document). At an early stage we made the pragmatic decision to encode manuscript descriptions in the header, thus treating them as descriptive metadata rather than as source texts, in large part to allow the `<text>` element to be used for transcription of the manuscript itself. However, this decision to treat catalogue descriptions only as metadata *about* manuscripts rather than as texts to be transcribed (and described) in their own right is perhaps the source of some of the confusion which can surround the use of manuscript catalogues as datasets (the same would apply to the transformation of manuscript catalogues into MARC or EAD records, or into other metadata formats). 35

What gets lost is the concept of catalogues as texts in themselves, with their own contexts, histories, styles, and sources. By obscuring the complex nature of catalogues as texts in their own right, we invite their use as “always-already present unifying entit[ies]” with all of the accompanying problems and pitfalls that entails [Baker, Salway, and Roman 2022, 2]. It may be that in future projects the creation of digital editions of manuscript catalogues from which data about manuscripts can be derived as a secondary process would be more helpful in retaining a sense of the nature and provenance of the data that they contain, particularly when adding context to the results of computational processes. 36

6 Methodology and Results

6.1 Summary

Our core method was to group together manuscripts that share selected physical features and then to examine those groups to see if information derived from manuscripts with an established place of origin could be extrapolated across those for which this information is lacking. The scripts were written for use with the Fihrist dataset but should work on 37 any TEI dataset that uses the Manuscript Description module. The scripts are in Python and have been deliberately kept short and simple (no more than 200 lines of code) to facilitate reuse. The scripts have been documented and made openly available on GitHub.

The first stage in our methodology was to identify relevant manuscript features (such as script, language, size etc.) in consultation with subject specialists. We then extracted these features from the TEI dataset and normalised them to give values that would group effectively (see below for details). Quantitative analysis of general patterns in this subset of data (such as proportion of manuscripts originating from different countries, number of decorated manuscripts, etc.) gave us pointers as to which features were likely to be indicators of place of origin and should therefore be included in the grouping process. 38

The subset of normalised data was then grouped into sets of manuscripts that shared these features, and any grouping of 10 or more manuscripts was evaluated by the subject specialists to see whether the information on the place of origin was significant. At this stage, the subject specialists fed back into the grouping process, suggesting new or refined features, and the grouping was re-run until satisfactory results were obtained. The final stage was to integrate the newly-generated hypothetical information back into the Fihrist dataset. 39

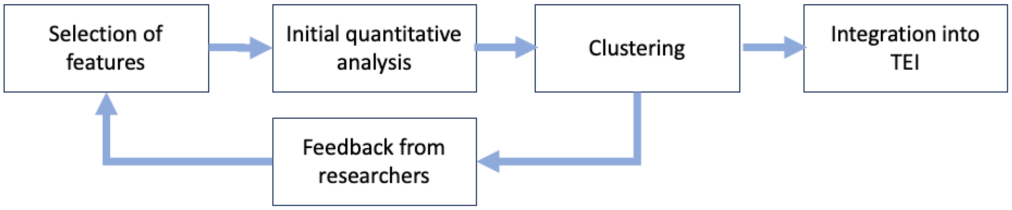


Figure 2. Project workflow.

6.2 Selection Criteria and Data Extraction

In consultation with subject specialists, we assembled an initial list of criteria that were a) recorded in TEI dataset in such a way as to be available to computational analysis, and b) considered most likely to provide clues as to the place of origin of the manuscripts. The criteria selected were: the material the manuscript was written on, the script, the language, whether or not the manuscript was decorated, the orientation, the number of written lines, and the size of the manuscript. 40

We then built a subset of TEI data containing these criteria, along with information allowing us to identify the manuscripts and the provenance information we were attempting to extrapolate across the groups: 41

- Criteria used to identify the manuscript: msid (system id), filename, repository, collection, idno (classmark/shelfmark)
- Criteria we wanted to extrapolate across the groups: start_century, end_century, origplace, country
- Criteria used in the grouping process: material, script, mainlang (language), decorated, orientation, lines, size

Some of these (msid, filename, repository, collection, idno, origplace, material, script, mainlang) were extracted directly from values in TEI elements or attributes. The remainder (start_century, end_century, country, decorated, orientation, lines, size) were derived from TEI elements and attributes but processed (binned) to give broader categories of values that were more suitable for grouping than the very specific values used in TEI: 42

- start_century and end_century: Derived from ISO dates recorded in the from, to, and when attributes on the `<origPlace>` element
- country: Automatically generated from the contents of the `<origPlace>` element using the GeoPy Python

library

- decorated: A binary yes/no value derived from the presence or absence of at least one `<decoNote>` element
- orientation: Landscape or portrait, derived from height and width values in the `<dimensions>` element
- lines: 1-10, 11-20, or 21_plus, derived from the `@ruledlines` attribute on the layout element
- size: L (large), M (medium), or S (small), derived from the height and width values in the `<dimensions>` element

43 The script that extracts and processes these values, **tei_to_csv_and_dataframe.py**, currently relies on having a copy of TEI dataset locally, but it could be easily reconfigured to run off an API (e.g., Cambridge Digital Library's TEI API, or, with some amendment to the code, IIIF manifests). It generates two outputs: a csv file, which is a versatile file for input into a range of mapping and network analysis tools, and a Python dataframe file, which is used as the input to the grouping process.

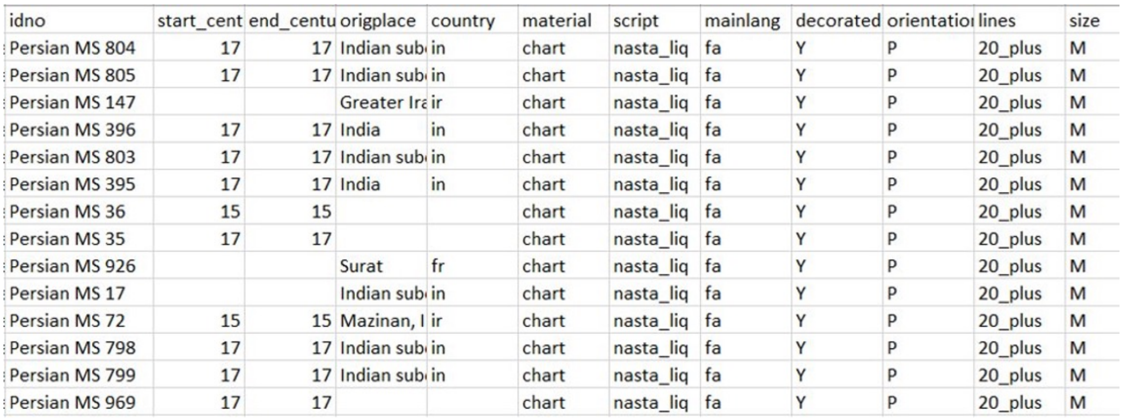
6.3 Initial Quantitative Analysis

44 Quantitative analysis of the data was performed using the **dataframe_to_plot.py** script. The primary purpose of this script was to establish general patterns in the data which would be helpful both in selecting criteria for grouping and in evaluating the grouping process. The script takes as input Python dataframe files and outputs graphs showing statistical information by time (century) or by place (country). You can also choose to view the statistics either by number of manuscripts with a particular feature, or by percentage of manuscripts with a particular feature.

6.4 Grouping

45 The subset of data in the Python dataframe file generated by **tei_to_csv_and_dataframe.py** was then grouped using the **dataframe_to_clusters.py** script. This script prompts you to pass in a list of features by which to group the data. By default, any groups containing fewer than 10 manuscripts are ignored (this threshold can be easily adjusted in the script). Those containing 10 or more manuscripts are output to separate, numbered csv files, one for each group.

46 The groups were assessed by subject specialists to determine a) if the information on place of origin was significant, and b) if there were patterns that would allow us to hypothesise on the origins of manuscripts which lacked a firm provenance.



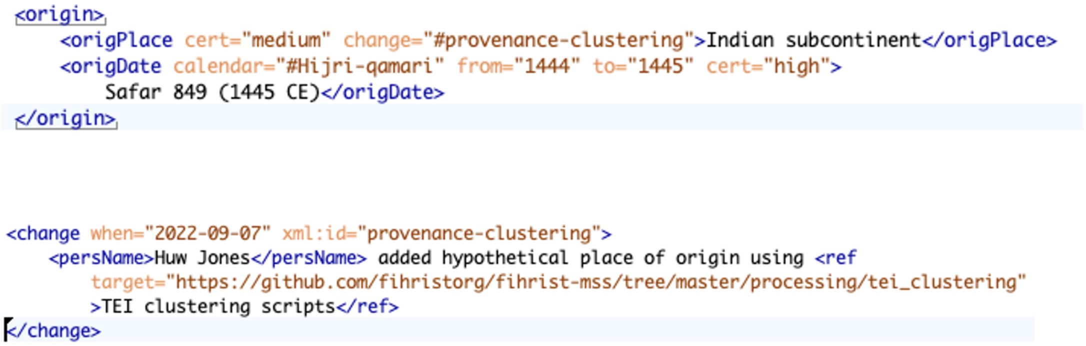
idno	start_cent	end_centu	origplace	country	material	script	mainlang	decorated	orientatio	lines	size
Persian MS 804	17	17	Indian sub	in	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 805	17	17	Indian sub	in	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 147			Greater Ira	ir	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 396	17	17	India	in	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 803	17	17	Indian sub	in	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 395	17	17	India	in	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 36	15	15			chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 35	17	17			chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 926			Surat	fr	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 17			Indian sub	in	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 72	15	15	Mazinan, I	ir	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 798	17	17	Indian sub	in	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 799	17	17	Indian sub	in	chart	nasta_liq	fa	Y	P	20_plus	M
Persian MS 969	17	17			chart	nasta_liq	fa	Y	P	20_plus	M

Figure 3. Example group showing the `origplace` value from the source data and the country value derived from it using the GeoPy library

6.5 Integration of Results into Dataset

47 Where we had established a hypothetical provenance for a previously unplaced manuscript, we integrated this data back into the TEI record. Here we used a combination of TEI's `@cert` (certainty) attribute and the `<revisionDesc>` element to a) indicate a level of certainty for our hypothesis, and b) reference both the method and the data on which

this hypothesis was founded. This stage was essential to one of the core aims of the project: to ensure that our own activities and methods were explicitly recorded in the dataset to inform future research. By default, TEI provides a limited range of values for `@cert`: “low”, “medium” and “high”. In this test project, we used the value “medium” by default, as we were not confident that the results of our processes could be classified as producing results with a high level of certainty, and where the correlation was tenuous we did not feed the results back into the dataset. It may be that a more granular approach to certainty would be necessary for a larger project with a broader range of results.



```
<origin>
    <origPlace cert="medium" change="#provenance-clustering">Indian subcontinent</origPlace>
    <origDate calendar="#Hijri-qamari" from="1444" to="1445" cert="high">
        Safar 849 (1445 CE)</origDate>
</origin>
```

```
<change when="2022-09-07" xml:id="provenance-clustering">
    <persName>Huw Jones</persName> added hypothetical place of origin using <ref
        target="https://github.com/fihristorg/fihrist-mss/tree/master/processing/tei_clustering"
        >TEI clustering scripts</ref>
</change>
```

Figure 4. Integrating results into TEI record with a `@cert` attribute to indicate level of certainty.

7 Case Study: Persian and Indian Manuscripts in the Collection of the University of Manchester

7.1 Collection and Cataloguing

Our test case was the University of Manchester collection, which was suitable both in terms of size (1,324 manuscript descriptions, so large enough for the testing of quantitative methods) and in terms of the richness of the descriptive records, having been the subject of a recent cataloguing project led by Jake Benson. The collection is also geographically diverse, “spanning the Balkans to the Bay of Bengal” [Benson 2021], and so allowed us to test our theories on the automatic generation of provenance information. 48

The collection also exhibited two other features general to Fihrist and to manuscript catalogues more widely. Firstly, the depth of description is not consistent across the collection, with manuscripts in Persian having richer records than manuscripts in Arabic and other languages. Secondly, cataloguing is a work in progress, based partly on new research and partly on legacy data derived from historical catalogues. 49

Also of interest is that the physical collection itself has a complex provenance, originating in large part from the Crawford Collection of Alexander Lindsay, 25th Earl of Crawford (1812–1880), with later additions from the collections of Samuel Robinson of Wilmslow (1794–1884) and of manuscripts formerly held at Manchester's Chetham Library. The Crawford Collection, itself the product of the acquisition of several preceding collections, had been the subject of previous cataloguing activities: a descriptive handlist drawn up by Michael Kerney and a list of titles, authors, and scribes published in 1898 [Benson 2021]. 50

7.2 General Quantitative Analysis

This dataset provided us with the opportunity to test the efficacy of our grouping method in first identifying and then distinguishing between manuscripts which come from two closely related traditions: Indian and Iranian. We would expect manuscripts from these traditions to share many features, such as script and language, but our proposition was that codicological information might allow us to hypothesise about which unplaced manuscripts belong to which tradition. 51

52 Of the 1,324 manuscript records in the dataset, 324 have a place of origin recorded. Of these, 190 are recorded as originating in India (59% of placed manuscripts) and 38 from Iran (12% of placed manuscripts). This might lead us to assume that of the 1,000 unplaced manuscripts, c.590 would be likely to be of Indian origin and c.120 of Iranian origin. However, it may, for example, be that the larger number of manuscripts recorded as being from India are not a reflection of general patterns in the collection but of a range of other factors, such as: a) it being easier to identify Indian manuscripts from textual or other data, b) an expertise in the identification of Indian manuscripts on behalf of the cataloguers who have described the collection, and/or c) extra attention being paid to the cataloguing of Indian manuscripts for funding or institutional reasons. While they do not help us directly in predicting the origin place of manuscripts, these overall statistical features of the dataset come in useful when assessing the significance of patterns which emerge in the grouping stage.

53 There are 66 manuscripts recorded as originating in Syria, but the catalogue information for these is too thin for them to effectively group — a common problem in the dataset as a whole, discussed below. There are a handful of manuscripts recorded as originating in other contexts: 4 from Afghanistan, 8 from Britain, 4 from Pakistan, 10 from Turkey, and 1 from Uzbekistan. It is worth noting that the Geolocation API's allocation of manuscripts with more general origin places (such as “Indian Subcontinent”) to India rather than to Pakistan highlights general issues with the translation of historical areas to modern nation states. The same issue applies to the allocation of manuscripts recorded as originating in the “Ottoman Empire” to Turkey — we address these issues and offer some possible strategies in Section 9.

54 Looking more closely at the data for Iranian and Indian manuscripts, information relating to text and language follows the pattern we would expect: 99% of Indian manuscripts are in Farsi (one manuscript is in Pashto), and 100% of Iranian manuscripts are in Farsi, reflecting the Mughal Empire's adoption of Persian as a state language. Script follows the same general pattern of similarity across the traditions:

Script	India	Iran
Naskh	9%	21%
Nasta-liq	86%	76%
Shikastah	3%	0%
No value	3%	2%

Table 1.

55 What was less encouraging for our specific research question was that the same patterns of similarity were reflected in the codicological information. 100% of both sets of manuscripts were on paper, making material (recorded at this level of granularity) essentially an irrelevant field in the grouping process. Almost all manuscripts across the traditions had a portrait orientation (95% in both cases). The traditions also exhibited very similar patterns in terms of the size of the manuscript and the number of lines per page:

Size	India	Iran
Small	16%	18%
Medium	71%	66%
Large	8%	11%
No value	5%	5%
Number of Lines	**India**	**Iran**
1-10	5%	2%
11-20	49%	50%
21 plus	27%	34%
No value	19%	14%

Table 2.

56 The only codicological feature where there was a notable difference between the two traditions was the number of manuscripts which were recorded as being decorated, with 68% of Iranian manuscripts recorded as decorated compared to only 29% of Indian manuscripts.

7.3 Grouping

57 With the exception of decoration, the individual codicological features, taken one by one, did not show notable differences between the two traditions. However, the grouping process gave us a chance to establish whether combinations of features (for instance small, decorated manuscripts with 11-20 lines per page) were more effective in distinguishing between manuscripts from India and Iran.

58 Informed by the general quantitative analysis, we ran the grouping script with 5 parameters: script, decorated, orientation, lines, and size (leaving out material and language which nearly always returned the same values for the two traditions). This produced 7 groups representing sets of 10 or more manuscript records which shared identical values for all of these features. Of the 1,324 records in the dataset, 169 (or 13%) were represented in the output groups — with other records either forming groups of fewer than 10 or lacking the necessary information to be included in the grouping process.

59 At first glance, the process seemed to be effective in identifying records relating to manuscripts which were from one of the two traditions (e.g., manuscripts that were *either* from India *or* Iran). Of the 169 records which formed groups of 10 or more, 136 had a place of origin recorded, with the majority from India (104) or Iran (25) and only a few from other countries (1 from Afghanistan, 1 from Pakistan, 3 from Turkey, and 1 from Uzbekistan), and 33 records lacking a place of origin. However, looking at cataloguing practices within the collection, it is possible that a major cause of this effect is the concentration of cataloguing activity on Persian and Indian material, leading to fuller records which are more likely to form into larger groupings.

60 Comparing the subset of records within the output groups to the whole dataset, there are slightly increased baseline probabilities of manuscripts being recorded as originating in India (62% up from 59%) and Iran (15% up from 12%).

61 Groups 1-7 contained manuscripts recorded as having the following features:

Group #	Script	Mainlang	Decorated	Orientation	Lines	Size
Group 1	nasta_liq	fa	N	P	11_20	M
Group 2	nasta_liq	fa	N	P	11_20	S
Group 3	nasta_liq	fa	N	P	21_plus	L
Group 4	nasta_liq	fa	N	P	11_20	S
Group 5	nasta_liq	fa	Y	P	11_20	M
Group 6	nasta_liq	fa	Y	P	21_plus	L
Group 7	nasta_liq	fa	Y	P	21_plus	M

Table 3.

The manuscripts in Groups 1-7 are broken down by origin place below:

Group #	Total Number of Records	Afghanistan	India	Iran	Pakistan	Turkey	Uzbekistan	No Place Recorded
Group 1	53	1 (2%)	40 (75%)	2 (4%)	1 (2%)	0 (0%)	0 (0%)	8 (17%)
Group 2	13	0 (0%)	7 (54%)	3 (23%)	0 (0%)	2 (15%)	0 (0%)	1 (8%)
Group 3	14	0 (0%)	12 (86%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (14%)
Group 4	19	1 (5%)	15 (79%)	0 (0%)	0 (0%)	1 (5%)	0 (0%)	2 (11%)
Group 5	40	0 (0%)	19 (48%)	10 (25%)	0 (0%)	1 (2%)	1 (2%)	9 (23%)
Group 6	14	0 (0%)	6 (43%)	7 (50%)	0 (0%)	1 (7%)	0 (0%)	0 (0%)
Group 7	15	0 (0%)	3 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (20%)

Table 4.

8 Assessment

62 Against a baseline background probability of manuscripts in groups being recorded as coming from India (62%) or Iran (15%), the groups do exhibit some clear tendencies. Manuscripts in Groups 1, 3, and 4 have a higher chance of being recorded as originating in India (with Group 3 at 86% being at first glance the most compelling example), and manuscripts in Groups 2, 5, 6, and 7 have a higher chance of being recorded as originating in Iran (Group 6 showing the strongest tendency at 50%). However, even for Groups 3 and 6, the size of the groups (14 manuscripts in each) makes it difficult to make a strong case for the generation of hypotheses on the place of origin of “unplaced” manuscripts within these groups. This highlights the central problem with our analysis: the number of catalogue records, even in a collection which has been the subject of a recent and comparatively large-scale cataloguing project, that have

sufficiently detailed codicological descriptions to form the basis of effective quantitative analysis.

Two other factors are clear in our test dataset. The first is inconsistency in cataloguing practices — the most obvious examples being the different terms for paper used in the records ("chart" and "paper") and the mixed use of both two- and three-letter language codes (e.g., "fa" and "per"). This is a wider problem in the description of manuscripts. Bridget Whearty, discussing similar variability in descriptors for parchment ("parchment", "parch", "membrane", "vellum"), references Ben Albritton's assertion that "… manuscript studies, as a field, is not interoperable with itself" [Whearty 2022, 171]. The second is the concentration of cataloguing efforts on certain parts of the collection, which then inevitably feature more heavily in the analysis — in this case, Iranian and Indian manuscripts. As a result, our test case is a microcosm of the larger Fihrist dataset, where, despite documentation of practices and the use of common standards, inconsistencies in the recording of values in the data are common, and there is enormous variability in the depth of cataloguing even within the collections of single institutions. 63

Even where rich data has been created, our efforts were hampered by a lack of granularity in the description of key codicological features. For Islamicate manuscripts, the classification of manuscripts as being written on "paper" (or "chart", see above) is so broad as to be meaningless in terms of analysis, with 100% of manuscripts in our test dataset exhibiting this feature. Other features, such as script or type of decoration would also benefit from more granular classification schemes, such as those suggested by Déroche and Gacek [Déroche 2006], [Gacek 2001]; see also the work on script classification done by [Prebor, Zhitomirsky-Geffet, and Miller 2020]. In some cases, such as the descriptions of inks, seals, and bindings, prose descriptions are not accompanied by formal classification schemes, making it almost impossible to include these features in computational analysis. 64

For some features, such as origin place (crucial to our research question) there was, perhaps inevitably, a mixture of very broad categories (e.g., "Indian Subcontinent") and more specific terms ("Delhi", "Lucknow", "Isfahan"), and the inclusion of values denoting historical political entities ("Mughal Empire", "Safavid Empire", "Ottoman Empire") alongside modern national classifications ("India", "Iran", "Turkey"). Issues regarding the derivation of meaningful entities for computational analysis from historic place names have been discussed at length in the context of the *Pelagios* project [Vitale et al. 2021]. While *Pelagios* has largely focussed on resolving place names as they appear in text, rather than places which we assert as the sites of origins of manuscripts, there is considerable overlap here as a number of origin places in our dataset are transcribed from colophons, stamps, or other text found in the manuscript. Our problems here were compounded by the lack of a comprehensive use of a formal identifier scheme for places in the Fihrist dataset. This led us to rely on automated conversion (using the GeoPy library) of the prose text in `<originPlace>` elements into codes relating to modern nation states for the generation of consistent units for analysis. In addition to a considerable error rate, this also raised the issue of the applicability of modern national boundaries to historical manuscript cultures. Some possible approaches to this problem are discussed in Section 9. 65

In common with the majority of catalogue datasets (and digital editions), Fihrist's current use of identifiers and vocabularies concentrates mainly on the intellectual contents of the manuscripts rather than on their history or physical manifestation. It makes extensive use of identifiers, both internal and external, for people and works — including VIAF identifiers where available. Machine-readable forms of dates are given, and there has been some effort to be consistent in the recording of language codes (ISO two-letter codes), the roles people have with regards to the manuscript (MARC relator codes), and material and script types (local schemes). However, one of the results of our analysis was to highlight considerable variance of practice even in these limited areas. In addition to the need for more granular vocabularies for material and scripts, as mentioned above, future work would benefit from similar controlled vocabularies or identifier schemes for the description of other codicological featuresk, with decoration, binding, inks, and seals being just a few examples. Put simply, if a feature is not consistently classified according to a controlled vocabulary, then even if it is described in detail it is almost impossible to include it in any kind of quantitative analysis, and if these vocabularies or identifier schemes are not standard across datasets, then the kind of large-scale comparative analysis promised by computational approaches becomes very difficult. 66

Finally, it may be that alternative methodologies would give more promising results with limited and/or inconsistent data. One option would be the application of multi-dimensional clustering to the dataset, which would give an idea of the 67 importance of specific features in establishing hypotheses about provenance and could lead to a more nuanced interpretation of the results of quantitative analysis based on the “weighting” of particular features.

9 Conclusions and Next Steps

There is no doubt that computational approaches to manuscript catalogues have the potential to “shed light”, though perhaps not on the things we might expect. The relationship between catalogues of manuscript collections and historical manuscript cultures is complex — and catalogue data has much to tell us about the processes by which collections of manuscripts come to us in their present forms and about the motivations and practices of those (including ourselves) who have created, described, and used those collections. The application of digital methods to catalogue data also highlights some important general issues in data-driven digital humanities. 68

In our project we used methods which went beyond simple inference or reference and looked at how patterns occurring in large datasets might tell us new things about manuscript collections. The technical aspects of the project did not raise significant issues — TEI is more than adequate as a data standard, the scripts “work” and do not in the first instance require complex coding in order to produce interesting results. Rather, the issues lie in the data. There is the practical problem of assembling datasets that are rich enough and are at sufficient scale to produce insights that would be impossible to obtain using traditional research methods. There is also the deeper issue of examining the practices and motivations of collectors and cataloguers in order to give context to the results of quantitative methods. 69

9.1 Strategies for Partial and Inconsistent Datasets

While many of the data issues which became clear during the project point towards the need for changes in cataloguing practices in response to the requirements of digital methodologies (see Section 9.2 below), there are some practical steps that could be taken with the current dataset to improve the results of quantitative analysis in the short term. 70

In the first instance, automated approaches to the identification and correction of inconsistencies in the data would produce a more reliable resource. One of the advantages of applying digital methodologies to datasets is that inconsistencies and errors of this kind tend to be highlighted at an early stage. Two examples that came up in the course of the project are the consistent use of two-letter language codes (e.g., “fa” rather than “per”) and attribute values for materials (e.g., “paper” rather than “chart”). Across the whole of the Fihrist dataset (c. 15,000 records), the identification and correction of inconsistencies arising either from mistakes in data entry or in variations in cataloguing practices is likely to make a significant difference in terms of the reliability of the results of quantitative analysis. 71

Another possible strategy would be the use of natural language processing to extract standard attribute values from the textual content of elements. The project identified a number of fields where standardised attribute values were present but were not granular enough to be useful in quantitative analysis. The most striking example was “paper”, which was so ubiquitous in the dataset as to be almost useless in terms of categorising material. Similar problems were encountered with the values used for script and decoration. The textual content of the elements often provides additional information (for instance, on the type of paper or sub-categorisation of script) which could be drawn upon to generate a more specific and granular attribute value. The same logic would apply to the generation of new attribute values where none currently exists for fields likely to give important information about the provenance of manuscripts, such as seals, inks, and bindings. However, the use of natural language processing to auto-generate values at scale needs to be approached with caution, as the free text content of elements may contain nuances that require an understanding of the context of specific terms within a sentence or paragraph (e.g., distinguishing between cases where a manuscript is *on* European paper from cases where European paper has been used for repairs). 72

One specific problem encountered in the project was the reconciliation of place names as recorded in the manuscripts with historical regions appropriate to our analysis. As described above, we used the GeoPy Python library to attempt to derive a broad idea of the regions of origin from the values recorded in the dataset. However, the returned values tended to reflect modern national boundaries, which were not always useful in the context of our research question. The building of project- or subject-specific gazetteers (along the lines of the *Pelagios* project) and/or the inclusion of suitable identifiers at the point of cataloguing are probably the best long-term strategies to address this problem (see Section 9.2 73 below), but for large sets of legacy data, automated reconciliation against existing resources could provide at least a partial solution. One possible route could be the Wikidata reconciliation functionality provided by tools such as OpenRefine, though we would then be relying both on the presence of the placenames in Wikidata and on the level of historical context provided in the entry. With similar caveats, automated reconciliation of values in the dataset against external resources could also be an option for other values in the dataset, such as scripts, seals, and materials. This would have the additional advantage of allowing us to automatically enrich our analyses with data from external sources.

9.2 Suggestions for Cataloguing Practices

In terms of strategies for future cataloguing practices, the project highlighted two core issues: the sheer labour involved in creating complex data at scale and the inherent connection between data, method, and research question. These are neatly summed up by Ezio Ornato (our emphasis): "… everything in a book is theoretically measurable, or, better yet, … each characteristic observed in a book may be formalized to extract measurable data from it, *provided it is worth the effort*. [Ornato 2020, 651]" 74

Which leaves us with the question: "*worth the effort*" to do what? There are an infinite number of things you can record about pretty much any object — what you choose to record (and how you choose to record it) depends on what you want to find out. Manuscript cataloguing is a type of data modelling, and, as Flanders and Jannidis say, " … the formalized model determines which aspects of the subject will be computable and in what form" [Flanders and Jannidis 2015, 229]. Given an infinite number of possibilities, deciding on what is "*worth the effort*" to record relies on an idea of the purpose of the model — what we want to "be computable and in what form". The blueprint of the model lies in the questions you want to ask. 75

Our pilot project — with a single research question based around provenance — identified a number of areas where the data would need to be enriched or improved significantly in order to produce useful results (and this in a collection which has been the subject of a recent and well-funded cataloguing effort). It is likely that other research questions would make similar demands of the data. We have some idea of the effort involved in producing the kind of very rich catalogue data that lends itself to quantitative analysis. In the Cambridge context, the recent Polonsky Foundation Greek Manuscripts Project produced 422 very detailed catalogue records through the efforts of three full-time cataloguers over three years, and the current Curious Cures in Cambridge Libraries Project plans to produce (amongst other things) just over 180 similarly detailed catalogue records, including partial transcription, in two years with three part-time research positions. This kind of output requires not only effort but also a broad range of expertise. What useful strategies can we adopt to embark on work of this scale? 76

We would advocate following the example of our pilot project in creating strong and explicit links between data creation and research questions. To go back to the "*provided it is worth the effort*" stipulation, all cataloguing activity is purpose driven, whether those purposes are explicit or not. By engaging in projects which are direct collaborations between researchers and cataloguers and therefore between data creation and research questions, we target our energies towards work that really is "*worth the effort*" and has a measurable way of assessing its value and efficacy. 77

But don't we run the risk of creating a dataset that is only responsive to certain uses? Perhaps, but there are practical measures which can mitigate against the risks of producing data that is specifically geared to only one set of research questions. Collaborative and comparative work of the kind facilitated by digital approaches could do much to usefully standardise manuscript catalogue data, particularly in areas such as codicological description where there are fewer large standard vocabularies or identifier schemes than in prosopographical or textual studies. 78

Our project identified several areas where there was a need either for more granular descriptive terms for physical aspects of manuscripts (material, scripts, decoration) or entirely new descriptive vocabularies (inks, seals, bindings). The development of standard vocabularies and/or identified schemes in these areas would not only address specific requirements arising from our project, but would also be more generally useful in addressing other research questions. The extent to which such vocabularies or identifiers would be able to span different manuscript cultures (Islamicate, Western European, Sanskrit, etc.) and create compatible datasets for interdisciplinary study is an interesting question 79 that raises important issues fundamental to the aims, methods, and purposes of manuscript description.

As discussed in Section 5.2, much of our data modelling work centres around our involvement in the collaborative development of a consolidated schema for manuscript description. The embedding of standard vocabularies arising from research projects into the schema would not only address the practical needs of projects engaged in quantitative analysis of manuscript descriptions, but also act as a forum for discussion of general issues arising from those projects — particularly the potential for quantitative analysis that spans different manuscript cultures and the kind of data models and standard vocabularies that would be required to embark on work of this type. 80

9.3 Manuscript Catalogues as the Focus of Research

While the automated correction and enrichment of datasets and changes to cataloguing practices would certainly improve the results of computational methods, our project highlights fundamental problems surrounding the correlation of the results of quantitative analysis with the realities of historical manuscript cultures. This raises the question: if quantitative approaches to manuscript catalogues are not well suited to establishing “hard facts” about historical manuscript cultures, then what are they good for? 81

To return to Ornato, one perspective on manuscript catalogues is as multi-layered representations of what people have thought it “*worth the effort*” to record about manuscript collections over time. If we aim to draw conclusions about the “realities” of historical manuscript culture directly from catalogue data, this can present a challenge. If, however, catalogues and their formation are framed as integral to our research questions, then they present an enormously useful resource for the understanding of changing attitudes towards manuscript collections, as well as serving as a mirror to contemporary social and political attitudes. 82

Our own work in the creation and curation of datasets derived from manuscript catalogues can be illuminating in this context. Many of the data issues raised by our pilot project centred around four core activities: selection, identification, authority, and method. We select which features we want to describe, we identify those features according to a typology or classification scheme, we indicate, explicitly or implicitly, by what authority we make this assertion, and this whole process happens in the context of method — current or anticipated use of the data. Viewing manuscript catalogues in this light has practical advantages in terms of understanding the results of quantitative methods by putting them in the context of the situations and motivations that produced them. In this we would argue against Drucker's distinction between the identification of “trends in the data” against “phenomena in the actual and lived world” [Drucker 2021, 115]. Data are phenomena of the actual and lived world and deserve our attention as such. A catalogue record for a manuscript is just as “real” as the manuscript itself. 83

Our project raised several potential research avenues in this area. In selecting our sample dataset, the inconsistency of depth and detail of description across the corpus became apparent — both between manuscript catalogues and within single catalogues themselves. Even in our sample dataset, there was a distinct difference between the level of information recorded about manuscripts written in Persian and other parts of the collection. An analysis of the relationship between types of material and level of detail recorded might produce interesting insights both at the micro and macro level, allowing us to trace the interests and specialities of individual cataloguers within catalogues and to analyse broader cultural and historical patterns of interest and engagement (including patterns in contemporary funding of collection description). 84

In addition to providing information on manuscripts, catalogues are also texts with their own historical and cultural context. The use of language in catalogues of Islamicate collections could be a valuable resource for allowing us to trace European attitudes to the Near and Middle East over time. A useful study could also be made of the type of person who was involved in the cataloguing of collections and by what “authority” (institutional, academic or cultural) they made their assertions about the manuscripts. Of particular interest here would be the identification of “lost” voices in the catalogues, including native speakers who made unacknowledged contributions to the work. 85

It would also be interesting to turn the focus back on ourselves and our own work. Decolonisation is often framed in terms of “fixing” historical biases in data, but it would be valuable to take a more reflective and critical stance on the 86 meaning and purposes of “decolonisation” as it relates to our project. Description and classification rely on sets of rules, and as Daston argues, “the island[s] of stability and predictability” created by rule-sets are “... the arduous and always fragile achievement of political will, technological infrastructure and internalized norms” [Daston 2022, 5]. Brown and Davis-Brown point to the political nature of this “achievement”, arguing that “...classifications never emerge solely from the material to be classified since our ways of defining the material itself are shaped by the dominant intellectual or political paradigms through which we view it” [Brown and Davis-Brown 1998, 25] — with our own classifications no exception. As we try to put the sources upon which our conclusions are based into their historical context, it is crucial that we acknowledge the “dominant intellectual or political paradigms” which shape our own activities.

9.4 Data Models, Digital Humanities, and Decolonisation

Data models and data modelling were central to many aspects of our project. The creation of the TEI dataset that formed the basis of our analysis was an act of data modelling, as was our own manipulation of the data in preparation for analysis. The legacy data that makes up much of the dataset has its own models, which encapsulate particular mindsets and attitudes and anticipate particular approaches. Data models in manuscript catalogues tend to divide collections into categories that are the embodiment of colonial classifications, such as the grouping of materials originating in diverse geographical contexts under the label “Oriental”. They also often claim authority over the content without referencing indigenous sources or contributions. Finally, they tend to anticipate that the study of different “manuscript cultures” will be undertaken separately, mitigating against interdisciplinary approaches across traditions. Digital data models often replicate these mindsets rather than challenging them. 87

We see an opportunity in the digital context to embark on data modelling activities that address decolonisation from a data-driven perspective. As we have seen, data modelling is not just about preparing content for publication and analysis; data modelling also frames what it is possible to do with the data. By using text encoding we can create descriptions that are based on conceptual models rather than strict hierarchies and formulations, and by working collaboratively across disciplines we can use this flexibility to create data that is inclusive and interoperable. We also advocate an open approach both to data availability and licensing which opens it up to audiences beyond the academic sphere — though there are real issues around the need for documentation, technology, skills, and infrastructure for new audiences to make use of the data. 88

In practical terms, this means moving away from traditional data models and towards structures, vocabularies, and identifiers that allow us to capture the individual characteristics of different manuscript traditions, while also facilitating interdisciplinary and comparative approaches. This is partly encapsulated in our collaborative work on the schema, but it also extends towards the establishment and use of vocabularies and identifier schemes — in particular a move towards taxonomies, which can group related phenomena together using numerical identifiers rather than relying on domain-specific language and terminology. A good example of the potential of these approaches was an analysis we undertook of the intellectual network of authorship of manuscripts from Cambridge's Islamicate collections and Oxford's Western medieval manuscripts (see Figure 5). This relied not only on these collections sharing a data model, but also on their use of VIAF identifiers, which allowed us to resolve the identity of authors who had been described in very different ways in the legacy data from which the datasets were constructed. 89

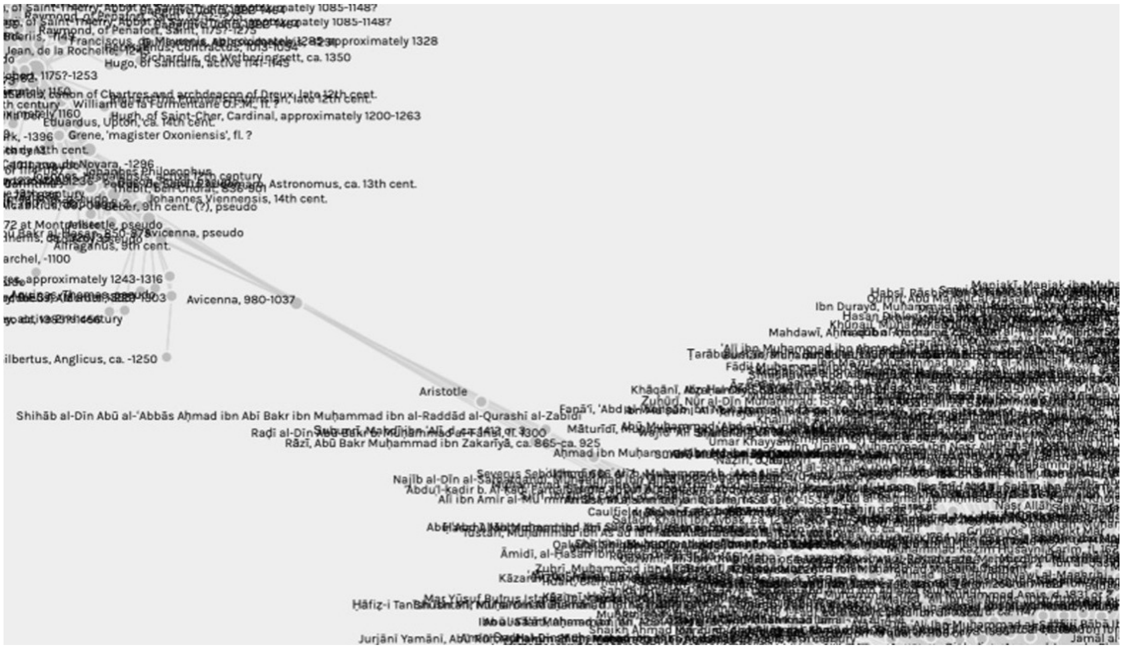

Raymond, of Peñafort, Saint, 1175?-1275
Richardus, de Wetheringsett, ca. 1350
Hugo, of Santalla, active 1141-1145
William de la Furmentarie O.F.M., fl. ?
Hugh, of Saint-Cher, Cardinal, approximately 1200-1263
Eduardus, Upton, ca. 14th cent.
Grene, 'magister Oxoniensis', fl. ?
Astronomus, ca. 13th cent.
Johannes Viennensis, 14th cent.
Avicenna, pseudo
Alfraganus, 9th cent.
approximately 1243-1316
Avicenna, 980-1037
ilbertus, Anglicus, ca. -1250
Aristotle
Shihāb al-Dīn Abū al-ʻAbbās Aḥmad ibn Abī Bakr ibn Muḥammad ibn al-Raddād al-Qurashī al-Zabīdī
Rāzī, Abū Bakr Muḥammad ibn Zakarīyā, ca. 865-ca. 925
Ibn Durayd
Mahdawī
Āmidī, al-Ḥasan ibn
Ḥāfiẓ-i Tanbūrī
Jurjānī Yamānī
Jamāl al

Figure 5. Network visualisation of authors in the Fihrist and Medieval Manuscripts in Oxford Libraries datasets, using Stanford University's Palladio software.

Finally, a note on data provenance. This project originated with questions about the origins of manuscripts, but the origins of data turned out to be equally if not more important. We have talked about the value of addressing the context in which legacy data was produced, and we also need to extend the same mechanisms to our own data modelling activities. Practically speaking, this means adopting a transparent and non-hierarchical model of recording contributions directly into the data, whether they be by people or by processes, and taking advantage of the versioning capabilities of modern repositories to record the various stages in the development of our datasets. As well as fostering a collaborative and iterative approach, this should mean that people using the data for similar projects in the future will have a more solid foundation for their analyses and a better idea of their point of departure. 90

Funding and Acknowledgements

This work was funded under the Cambridge Humanities Research Grants scheme. The Fihrist/OCIMCO catalogue was originally funded by JISC. Data cleanup and enrichment preparatory to the work was funded by Cambridge Digital Humanities. The Fihrist platform is hosted and maintained by the Bodleian Libraries, Oxford. 91

Notes

[1] For instance, if a scribe associated with a manuscript has attested dates in an external authority system such as VIAF (the Virtual International Authority File), then it is inferred that the manuscript must have been copied in that date range.

[2] For example, in a simple sense, if a scribe's dates are known then the manuscripts copied by that scribe must fall within that date range, or in a more complex sense, if a dated manuscript is associated with a scribe, then other manuscripts copied by that scribe must also exist in some relation to that date.

Works Cited

Baker, Salway, and Roman 2022 Baker, J., Salway, A., and Roman, C. (2022) “Detecting and characterising transmission from legacy collection catalogues”, *Digital Humanities Quarterly*, 16(2). Available at: https://www.digitalhumanities.org/dhq/vol/16/2/000615/000615.html.

Beit-Arié 2020 Beit-Arié, M. (2020) “Comparative codicology”, in Coulson, F. and Babcock, R. (eds.), *The Oxford handbook of Latin palaeography*. Oxford: Oxford University Press, pp. 669-673.

Benson 2021 Benson, J. (2021) “Cataloguing Persian manuscripts at the John Rylands Research Institute and Library”, *Rylands Blog*, 17 September. Available at: https://rylandscollections.com/2021/09/17/cataloguing-persian-manuscripts-at-the-john-rylands-research-institute-and-library/.

Brown and Davis-Brown 1998 Brown R.H. and Davis-Brown, B. (1998) “The making of memory: The politics of archives, libraries and museums in the construction of national consciousness”, *History of the Human Sciences*, 11(4), pp. 17–32. https://doi.org/10.1177/095269519801100402.

Burrows 2018 Burrows, T. (2018) “Connecting medieval and Renaissance manuscript collections”, *Open Library of Humanities*, 4(2), pp. 1-32. http://doi.org/10.16995/olh.269.

Burrows et al. 2021 Burrows, T., Holford, M., Morrison, A., Page, K., and Velios, A. (2021) “Transforming TEI manuscript descriptions into RDF graphs”, *Schriften Des Instituts Für Dokumentologie Und Editorik*, 15, pp. 143-154. Available at: https://ualresearchonline.arts.ac.uk/id/eprint/17511/1/graphsde2019_Burrows.pdf.

Chandna et al. 2016 Chandna, S., Rindone, F., Dachsbacher, C., and Stotzka, R. (2016) “Quantitative exploration of large medieval manuscripts data for the codicological research”, *IEEE 6th symposium on large data analysis and visualization (LDAV)*, pp. 20-28. http://doi.org/10.1109/LDAV.2016.7874306.

Cummings 2019 Cummings, J. (2019) “A world of difference: Myths and misconceptions about the TEI”, *Digital Scholarship in the Humanities*, 34(Supplement 1), pp. 58–79. https://doi.org/10.1093/llc/fqy071.

Daston 2022 Daston, L. (2022) *Rules: A short history of what we live by*. Princeton, NJ: Princeton University Press.

Daston and Galison 2007 Daston, L. and Galison, P. (eds.) (2007) *Objectivity*. New York: Zone Books.

Drucker 2021 Drucker, J. (2021) *The digital humanities coursebook*. New York: Routledge.

Déroche 2006 Déroche, F. (2006) *Islamic codicology*. London: Al-Furqān Islamic Heritage Foundation.

Faghihi, Holford, and Jones 2022 Faghihi, Y., Holford, M., and Jones, H. (2022) “Teaching the Text Encoding Initiative”, *Journal of Open Humanities Data*, 8. https://doi.org/10.5334/johd.72.

Flanders and Jannidis 2015 Flanders, J. and Jannidis, F. (2015) “Data modeling”, in Schreibman, S., Siemens, R., and Unsworth, J. (eds.) *A new companion to digital humanities*. London: Wiley, pp. 229-237.

Formigatti 2017 Formigatti, C.A. (2017) “Sanskrit manuscripts in the Cambridge University Library: Three centuries of history and preservation”, in Vergiani, V., Cuneo, D., and Formigatti, C. (eds.) *Indic manuscript cultures through the ages: Material, textual, and historical investigations*. Boston, MA: De Gruyter, pp. 3-46.

Gacek 2001 Gacek, A. (2001) *The Arabic manuscript tradition: A glossary of technical terms and bibliography*. Boston, MA: Brill.

Kinitz and Efer 2023 Kinitz, D. and Efer, T. (2023) “Towards a dynamic knowledge graph of a non-Western book tradition”, in Baillot, A. et al. (eds.) *Digital humanities 2023: Book of abstracts*. Alliance of Digital Humanities Organizations, pp. 216-217. Available at: https://zenodo.org/records/7961822.

Maniaci 2022 Maniaci, M. (ed.) (2022) *Trends in statistical codicology*. Boston, MA: De Gruyter.

Mylonas and Renear 1999 Mylonas, E. and Renear, A. (1999) “The Text Encoding Initiative at 10: Not just an interchange format anymore – but a new research community”, *Computers and the Humanities*, 33, pp. 1–9. https://doi.org/10.1023/A:1001832310939.

Ornato 2020 Ornato, E. (2020) “The application of quantitative methods to the history of the book”, in Coulson, F. and Babcock, R. (eds.) *The Oxford handbook of Latin palaeography*. Oxford: Oxford University Press, pp. 651-668.

Prebor, Zhitomirsky-Geffet, and Miller 2020 Prebor, G., Zhitomirsky-Geffet, M., and Miller, Y. (2020) “A new analytic framework for prediction of migration patterns and locations of historical manuscripts based on their script types”, *Digital Scholarship in the Humanities*, 35(2), pp. 441–458. https://doi.org/10.1093/llc/fqz038.

Salway and Baker 2020 Salway, A., and Baker, J. (2020) “Investigating curatorial voice with corpus linguistic techniques: The case of Dorothy George and applications in museological practice”, *Museum and Society*, 18. https://doi.org/10.29311/mas.v18i2.3175.

Steedman 1998 Steedman, C. (1998) “The space of memory: In an archive”, *History of the human sciences*, 11(4), pp. 65–83. https://doi.org/10.1177/095269519801100405.

Van Lit 2019 Van Lit, L.W.C. (2019) *Among digitized manuscripts*. Leiden, Netherlands: Brill.

Vitale et al. 2021 Vitale, V., de Soto, P., Simon, R., Barker, E., Isaksen, L., and Kahn, R. (2021) “Pelagios: Connecting histories of place”, *International Journal of Humanities and Arts Computing*, 15(1-2), pp. 5-32. Available at: https://www.euppublishing.com/doi/full/10.3366/ijhac.2021.0260.

Whearty 2022 Whearty, B. (2022) “Interoperable metadata and failing towards the future”, in *Digital codicology: Medieval books and modern labor*. Redwood City, CA: Stanford University Press, pp. 168-211.

Zaagsma 2022 Zaagsma, G. (2022) “Digital history and the politics of digitization””, *Digital Scholarship in the Humanities*, 38(2), pp. 830–851. https://doi.org/10.1093/llc/fqac050.

Zhitomirsky-Geffet, Prebor, and Miller 2020 Zhitomirsky-Geffet, M., Prebor, G., and Miller, I. (2020) “Ontology-based analysis of the large collection of historical Hebrew manuscripts”, *Digital Scholarship in the Humanities*, 35(2), pp. 688–719. https://doi.org/10.1093/llc/fqz058.