Information Research, Special Issue: Proceedings of the 15th ISIC - The Information Behaviour Conference, Aalborg, Denmark, August 26-29, 2024

Analysing humanities scholars' data seeking behaviour patterns using Ellis' model

Wenqi Li, Pengyi Zhang, and Jun Wang

DOI: https://doi.org/10.47989/ir292835

Abstract

Introduction. Much of humanities data are existing materials. Yet there are few works examining humanities scholars' data seeking behaviours. The study aims to address this gap building upon Ellis' model.

Method. We recruited 27 humanities scholars with diverse academic backgrounds and conducted in-depth interviews.

Analysis. A preliminary codebook was developed from existing literature. Researchers combined deductive and inductive coding to analyse the interview transcripts.

Results. Humanities scholars' data interactions fall into two approaches: data-driven and structure-driven, each involving three phases – exploratory seeking, focused seeking, and supplementary seeking. We identified eleven characteristics of data seeking behaviours operating at different levels and revealed their variations across research approaches and seeking phases.

Conclusion. The study contributes to the conceptual growth of Ellis' model and expands its utility beyond the original information seeking contexts, indicating its potential applicability to data seeking. It also provides practical implications for system design and humanities data curation.

Introduction

An increasing number of humanities studies are embracing a more data-driven approach (Borgman, 2010; Schroeder, 2014), expanding their research materials from traditional sources to encompass diverse forms of data, including texts, images, videos, databases, artefacts, and even algorithms (Flanders and Muñoz, 2012; Moulaison-Sandy and Wenzel, 2023). Consequently, new forms of humanities scholarly activities are emerging, extending beyond information seeking and use, but also data collection, processing, and analysis (Anderson et al., 2010; Pacheco, 2022; Palmer et al., 2009).

Research starts to pay attention to the data practice in the humanities (Borgman, 2015; Hoekstra and Koolen, 2019; Late and Kumpulainen, 2022; Ma and Xiao, 2020). Yet, more individual-oriented and nuanced understandings of the behavioural characteristics are needed to inform the design and implementation of data related tools, services, and infrastructure. In science and social science disciplines, quite a few studies have investigated scholars' data seeking, sharing, and reuse behaviours (Gregory et al., 2019; Gregory et al., 2020; Rolland and Lee, 2013; Yoon, 2017). While disciplinary approaches shape the way scholars interact with data (Chao et al., 2015), data behaviours of humanities scholars and their variations by research approaches require further examination.

Different from sciences and social sciences, much of the humanities data are not specifically generated for research purposes. Humanities scholars often rely on existing materials created by others or curated as part of humanities research infrastructure (Borgman, 2012; Schöch, 2013; Trace and Karadkar, 2017). Thus, seeking for data composes a significant portion of data collection, and is the foundational step of humanities studies. However, there is little research that investigated humanities scholars' data seeking behaviours.

Building upon the definition of information seeking behaviour (Wilson, 2000), data seeking can be conceptualised as the purposeful seeking of existing data or sources to satisfy specific research goals. It complements the generation of new data through field or lab research within the broader process of data collection (Chao et al., 2015). Given the fact that documents and publications are the most crucial type of humanities data (Gualandi et al., 2022), the boundary between data and information is often blurred (Borgman, 2008). Data seeking distinguishes itself from information seeking by its target on further analysis to extract meaning or insight. Further, the tasks and activities involved in data behaviours have similarities with information behaviours, such as need identification, retrieval, selection, and use (Rolland and Lee, 2013; Wang et al., 2021; Zimmerman, 2007). Thus, theories and models of information seeking behaviour may inform the investigation on data seeking behaviours. Particularly, Ellis's model examined academic information seeking behaviours in detail. Several extended models concentrated on humanities scholars (Ellis, 1989a; Savolainen, 2017), and some of them have benefitted from understanding seeking characteristics by phases (Bronstein, 2007; Rhee, 2012).

Therefore, this paper aims to explore the nuances of humanities scholars' data seeking behaviours, building upon Ellis' model while drawing from data behaviour investigations in other disciplines. The paper answers two research questions: (1) What are the characteristics of humanities scholars' data seeking behaviours? (2) How do these characteristics vary by research approaches or data seeking phases?

Literature review

Data practices in the humanities

Borgman (2015) examined data scholarship in the humanities with two case studies, highlighting the importance of data provenance and data representation. Ma and Xiao (2020) revealed that archival and bibliographical research, digitisation, and extracting data from databases are the most common data collection methods in digital history research. Hoekstra and Koolen (2019) proposed the concept of data scopes to demonstrate the iterative data transformation process in historical research. They argue that data interactions should be viewed as an integral part of doing research. Oberbichler et al. (2022) constructed an interdisciplinary digital interpretive research workflow starting with selecting and digitising the source materials. Late and Kumpulainen (2022) summarised information interactions of humanities scholars when using digitised newspapers for research, including various data behaviour activities, such as acquiring access to data in the task planning stage, analysing metadata when working with items, and opening the data during the synthesising and reporting stage. Koolen et al. (2020) proposed a hierarchical model for humanities research, integrating research stages, tasks, and information or data activities. They also suggest that users combine multiple sources and activities such as retrieving, browsing, and extracting to make sense of data. These studies provided a general understanding of data practices in humanities. Yet, there's lack of deeper investigation of humanities scholars' data seeking behaviours.

Whitmore's (2016) work is more relevant in that it analysed how scholars seek and process spatial information in archaeological research. But the study focused more on the purpose of seeking, sources and resources used, and accessing constraints, while lacking in-depth analysis of the behavioural characteristics.

Humanities scholars' information seeking behaviours and Ellis's model

Earlier studies have explored humanities scholars' information seeking in libraries, archives, and digital environments (Al Shboul and Abrizah, 2016; Buchanan et al., 2005; Duff and Johnson, 2002; Wiberley and Jones, 1989). For humanities scholars, search cannot replace browsing, although browsing in the digital environment can be difficult (Buchanan et al., 2005). Humanities scholars centres on core primary sources and takes a centrifugal path to seek for information sources (Palmer, 2005), relying on chaining and access tools to identify and locate materials and build contextual knowledge (Duff and Johnson, 2002; Palmer, 2005).

Ellis's model provided detailed characteristics of information seeking behaviours in academic contexts, including starting, chaining, browsing, extracting, monitoring, and differentiating (Ellis, 1989a, 1989b). The model was then extended by a few scholars, several of which investigated humanities scholars' information seeking behaviours (Bronstein, 2007; Ge, 2010; Rhee, 2012; Smith, 1988).

Wilson's expansion on Ellis's model made the earliest attempt on restructuring the components, integrating Kuhlthau's ISP model to arrange the information seeking characteristics into a staged linear process (Wilson, 1999). Bronstein (2007) further introduced the role of research phase to structure the seeking characteristics into initial phase and current awareness phase, and she also identified information managing and evaluation activities to be phase-independent elements. In fact, many studies have incorporated research phases in examining information behaviours (Brown, 2002; Chu, 1999; Koolen et al., 2020; Kuhlthau, 1991; Late and Kumpulainen, 2022). Data seeking behaviour investigations could also benefit from such a staged view (Wang et al., 2021).

Makri et al. (2008) made systematic conceptual elaborations on Ellis' model (Savolainen, 2017). The model inherits Ellis's view that there should be no sequential relationships among the elements. Instead, the various characteristics (lower-level behaviours) are categorised into three higher-level behaviours: identification and locating, accessing, and selecting and processing. Additionally, the model provides a granular elaboration with the levels that each behaviour is observed to operate at – resource, source, document, and content levels. These levels provide a directly actionable dimension to examine the information behaviours. Just as Anderson et al. (2010) suggested, analysis of scholarly activities should be directly informative to technology and data service providers.

Methodology

This study is part of a larger investigation into the data behaviours of humanities scholars. We conducted in-depth interviews to gain a comprehensive understanding on how they interact with data throughout the research process, including data sources and data seeking behaviours. The interview guide is included in Appendix I.

We adopted a purposeful sampling strategy to recruit 27 participants (12 females and 15 males) with diverse academic backgrounds, including different ranks, disciplines, and familiarity levels with digital humanities. We started with convenience sampling and expanded through referrals to include a broad spectrum of humanities scholars, aiming to capture diverse research approaches and data behaviours. The interviews were conducted from April to September 2023. Table 1 summarises the academic backgrounds of the participants.

Academic Background		# of Participants
Academic Ranks	Master's student (M1-4)	4
	PhD student (D1-11)	11
	Post-docs (P1-2)	2
	Faculties (F1-10)	10
Disciplines	History	8
	Language	5
	Philosophy	5
	Philology	4
	Literature	3
	Arts	2
Digital Humanities Familiarity	Familiar	17
	Unfamiliar	10

Table 1. Academic Backgrounds of Participants

We conducted online or face-to-face interviews based on participants preferences. Prior to each interview, we obtained written consent from participants, ensuring they were fully informed about the study's purpose, the confidentiality of their responses, and their rights to withdraw at any time. The average length of interviews is 85 minutes, ranging from 45 minutes to 2 hours. The interviews followed a pre-defined protocol, and we probed deeper as need. We recorded and transcribed the interviews for analysis.

During interviews, we asked participants to share or demonstrate any mentioned data-related artefacts. This process helped clarify and triangulate the interview data and facilitated deeper discussions. Most participants provided demonstrations, and 10 shared relevant screenshots or photos after the interviews. These artefacts were integrated into the transcript at corresponding points for further coding and analysis.

To analyse the data, we developed a codebook referring to existing framework (Bronstein, 2007; Chao et al., 2015; Ellis, 1989a; Ellis et al., 1993; Ellis and Haugan, 1997; Gregory et al., 2019; Makri et al., 2008). Combining deductive and inductive coding, two researchers coded six transcripts using the initial codebook, while also allowing for the emergence of new codes. The research team discussed coding results and refined the codebook through an interactive process. The two researchers then recoded the same six transcripts using the final codebook (see Appendix II), achieving a substantial inter-coder agreement (Cohen's Kappa=0.71). After resolving disagreements, they proceeded to code the remaining transcripts separately.

Findings

Phases of data seeking in humanities research

According to participants' description of their overall research process and interactions with research materials, two themes emerged during the initial coding process, categorising scholars' data behaviours into two approaches - the structure-driven and data-driven. Akin to the two types of sensemaking mechanisms, the data-driven research approach refers to the process of discovering patterns or forming theories through inductive analysis based on a substantial amount of data or texts. The structure-driven approach, on the other hand, begins with constructing a discourse structure based on initial reading accumulation, then fitting data to the structure to support the argument, and ultimately forming humanities observations or interpretations. Participants show a strong inclination towards one approach or another. 13 participants adopt a data-driven approach and 14 follow a structure-driven approach, with two from the latter group considering the data-driven approach for future research.

In both approaches, we identified three phases of data seeking from the coding: the *exploratory seeking* phase that is prior to the identification of research questions and information needs; the *focused seeking* phase that is driven by established information needs; the *supplementary seeking* phase that takes place during data processing and analysis.

In data-driven research, the role of data is similar as in social sciences – the evidence to be processed and analysed to form or validate a theory. Some participants mentioned that the data is also their research object. For example,

> *These three ancient books are my data as well as my research objects. What I study are the patterns in them. The research objects and data are basically united, which means that data is derived from the sources or so-called research objects.* (Participant P1 in ancient Chinese)

Therefore, the boundary between information and data is clear to data-driven scholars, and scholars are well-aware whether they are seeking data or information. Before initiating a study, scholars may exploratorily seek data to discover potential topics, or accumulate relevant data to update their personal collection. Once the research questions and data needs are identified, scholars begin focused data seeking. In this phase, data-driven scholars seek data extensively based on their needs and feasibility. In most cases, they want the data to be comprehensive or at least can represent the totality. They then process and analyse the data to discover patterns or build theories and begin writing concurrently. During this process, scholars may remain vigilant for potentially overlooked data or data previously inaccessible to them, which is the supplementary data seeking phase. The data added in this phase is unlikely to overturn their previous findings but mainly instantiate the structure, given the thoroughness of their focused seeking.

> *I initially try to be systematic (with data seeking), but then I just keep adding as new things come up, and there's quite a lot to add because something is always missed. At first, it was frustrating... but eventually, I just accepted it. Once your research is complete, you realise that the data you added later wouldn't significantly alter the structure of your original work. For instance, if you had prepared 10 images of a certain type and then found a new one, increasing it to 11 images, it usually doesn't lead to major changes.* (Data-driven participant F1 in art history)

In structure-driven research, the boundary between data and information is more obscure, thus it's difficult to discern data seeking from information seeking. Participants generally seek for documents or sources, which only become data if they are used as (1) objects of interpretation, rebuttal, or scholarly discourse; (2) evidence, example, or quotation to make an argument. During exploratory seeking, their primary purpose is developing their personal collection that could become data in their future research. When initiating a new study, they begin focused data seeking. Scholars narrow down to core ideas after preliminary search and radiate outward from the core material or author. For example,

> *I usually start with broad searches using a few key words, then narrowing down and selecting from what I found. I've taken lots of notes. But when I start a new writing, I*

> *continually condense these notes – from over a hundred pages down to maybe 10 or 5, focusing on core ideas… Initially, I read broadly like casting a wide net, but now I'm more focused, specifically on the asymmetry of nature and entities. I'm more targeted, tracking a few scholars closely and focusing on their works.* (Structure-driven participant D11 in western philosophy)

Unlike data-driven scholars, structure-driven scholars do not aim for exhaustiveness in focused seeking, but target on acquiring key sources to start interpretation and developing core arguments and discourse structure. Once initially establishing the framework, they begin writing, fitting their data to the structure, in forms of evidence, examples, or quotations. During this process, new data needs may arise, prompting scholars to seek additional data, marking the phase of supplementary data seeking. Comparing to data-driven scholars, structure-driven scholars persist in active data seeking during this phase. The new data acquired can either initiate the structure or alter existing arguments, leading to re-structuring. Although structure-driven scholars do not pursue large scale of data, their research is still based on substantial data. They conclude their data seeking when they believe they have gathered sufficient evidence to support their arguments and the discourse is coherent.

Humanities scholars' data seeking characteristics

Model overview

Based on the coding results, we identified eleven data seeking characteristics operating at different levels – the resource level, document or dataset level, and content level. We further mapped them to corresponding research approaches and seeking phases (see Table 2).

The data seeking characteristics include browsing, searching, networking, starter reference, chaining, selecting, accessing, verifying, extracting, representing, monitoring. All of them were covered in previous Ellis' model and extensions, except for representing. Detailed definitions of these characteristics and operation levels can be found in Appendix II.

The observed characteristics vary across different seeking phases as well as between data-driven (DD) and structure-driven (SD) research approaches. As shown in Table 2, the exploratory and focused seeking phases exhibit much similarity in characteristics between data-driven and structure-driven approaches. The characteristics unique to the research approach in exploratory and focused seeking phases are marked in the table. In the supplementary seeking phase, however, the difference between two approaches is more pronounced. Data-driven scholars primarily follow certain resources to monitor additional data, while structure-driven scholars perform similar activities as in focused seeking phase.

	Exploratory	Focused	Supplementary	
			Data-driven	*Structure-driven*
Resource Level	Browsing Searching	Searching Browsing (*SD*) Accessing Selecting (*DD*)	Monitoring	Searching Accessing
Document / Dataset Level	Networking Representing (*SD*)	Starter reference Chaining Selecting Accessing Browsing Representing (*SD*)		Chaining Selecting Accessing Browsing Representing
Content Level	Extracting	Extracting Verifying		Extracting

Table 2. The humanities scholars' data seeking characteristics operating at varying levels across different seeking phases

The following sections will explain the characteristics in detail, following the sequence of identifying and locating the data; the significance of chaining; acquiring the data; ensuring data quality; and occasional monitoring.

Searching, browsing, and extracting

These characteristics are primarily for scholars to identify and locate documents at resource level or extract the data they need at document and content level.

Both data-driven and structure-driven scholars rely heavily on searching online resources, such as databases, corpus, and online catalogues or finding aids of libraries and archives. Searching takes place in both exploratory and focused seeking phases, and structure-driven scholars' supplementary seeking phase. Most participants in our study referred to searching as *keyword searching*. Searching at resource level is often metadata search, as few databases provides full-text searching, especially for the dated documents that are not fully digitised. For example,

> *They (databases) have scanned these newspapers. Then, they manually capture the title and the author, inputting the date and journal number, allowing you to search by these fields. However, they haven't been able to digitize the entire article. It's still a photocopy, not fully digitized. For instance, it's not possible for me to search for articles that mentioned 'European war'.* (Structure-driven participant D3 in modern Chinese history)

Though searching provides immediate findability for the data needed, it may not provide a holistic view of the bigger contexts that are important to humanities studies. Participant M3 in contemporary literature thinks '*searching the database feels mechanical, fragmented, and turned the way of literary studies upside down*'. Yet due to limited time and the convenience of online resources, searching is still participants' first choice over browsing.

However, for structure-driven scholars, browsing the resource is irreplaceable when the sources they need are not well digitised, such as archives and local gazetteers. For example, D4 in modern history mentioned, '*I first go through the online finding aids of the archive to get an idea of which documents I need. Then, I go there and look for these documents.*' For data-driven scholars, browsing at the resource level only appears in the exploratory phase of data collection for data-driven scholars, because they mostly use digitised documents or dataset as data and can search for data online or through internal team sharing. In the exploratory searching phase, browsing can lead to emergence of ideas. For example, '*In film history, you might find interesting topics just by randomly picking a year's newspapers and browsing them. For me, I found an intriguing topic on the first day I accessed and browsed a newspaper database*'. (Data-driven participant F2 in film history)

Meanwhile, browsing at document level is also necessary for both structure-driven and data-driven scholars to make selecting decisions or extract the content they need, especially when digitised full text is not available. For example, participant D4 said he needed to '*sift through the publicly published archival compilations or dossiers.*' And data-driven participant F2 mentioned, '*Local gazetteers are not as digitized as newspapers and journals. I had to locate physical copies of these gazetteers and flip through them page by page.*' Browsing at document level may also lead to extracting the content. For example,

> *I already knew which part of the archaeological report contains data about the tomb, which part is about the artefacts, and which part details the bronze items. You get a general idea of its framework without reading through each section in detail. Just look for the data you need and record it once you find it.* (Data-driven participant D1 in ancient Chinese history)

For data-driven scholars, extracting can also be done other than browsing and manual extraction. In well digitised documents or data sets, they can use keywords, existing dictionaries, or running regular expression queries to automatically locate and extract data they need. Another layer of extracting is to process the contents to structured data, which is a more complex data processing behaviour that's not within the scope of this paper on data seeking.

Starter reference and chaining

Chaining is important for both structure-driven and data-driven scholars to locate more data. Besides forward and backward citation tracking, starter reference is also very important for humanities scholars to start their chaining process. The starter reference is usually a monograph by prominent scholars in the field, or a well-known reference book of the subject, such as bibliographies, image catalogues, indexes, and dossiers. For structure-driven scholars, it is just a map to locate more data. They describe the chaining process as '*looking for fine horses using only a picture*', '*crime detection*', or '*spin silk from cocoons*'. For example,

> *There are some prominent scholars who will provide a sort of topographical map that outlines the important documents in our field. So, we can follow the map to trace the important documents he mentioned. It's what we call archaeology of knowledge. That is, you may get a thread, then you follow the thread deeper to see where it is in the text chain. These documents may be referred in a monograph or included in some bibliographies. You just need to pull them out one by one.* (Structure-driven participant M3 in contemporary literature)

For data-driven scholars, the starter reference can also be where scholars extract contents to form their own dataset and then become their data. For example, participant D1 studying history formed her preliminary dataset based on a dossier of bronze inscriptions and an index of clan-sign inscriptions. The identification number in both reference books were unified, which was directly used as the URI in her dataset. Participant M1 studying modern Chinese used the dictionary 800 *Words of Modern Chinese* to derive the initial directional word list, based on which she formed the corpus for discovering the usage patterns. Thus, both the dictionary and the corpus were considered her data.

Networking, accessing, and representing

These three characteristics are all closely related to how scholars acquire the data for research use after identifying and locating them. In the exploratory seeking phase, both structure-driven or data-driven participants may obtain some documents or datasets through networking, which are relevant to their field of interests and at some points would be used as data. For example, data-driven participant F8 in ancient philology mentioned, '*I definitely have accumulated some texts* *relevant to my research focus, some of which were kindly given by teachers or senior colleagues and classmates at the time.'*

In the focused seeking phase, however, networking is seen as a part of accessing. Informal personal requests or collaborative sharing are both important means of gaining access at document level. Particularly, the digitised documents or datasets used by data-driven scholars are often shared and circulated within teams, which presents fewer barriers for accessing at document level.

However, accessing at resource level witnesses more constraints including cost, geography, and permission. For example, university libraries only grant permissions to their own students, rare book collections are even harder to access. Local archives are not transparent to the public due to political issues or require complex procedures for people to get in.

Representing is a characteristic that was uncovered in previous studies. It is closely related to the nature of structure-driven humanities studies, where a lot of the retrospective documents are not digitalised or digitised. Representing to some point is a way for them to bypass the access hurdle. For example, many structure-driven participants scan books that they could not borrow from the library or ask friends to scan them due to geographical or permission constraints. Participant D3 studying contemporary intellectual history mentioned that the data she needed is in an appendix of a book only preserved in national library. While photocopy services are expensive, she took pictures of all the appendices. And some local archives don't even allow digital products, so that she could only take notes or transcribe the archives.

Further, digitalisation or digitisation is also critical for scholars' long-term preservation and convenient access of their documents. For example,

> *I purchased these books oversea and carried them back to my dormitory. They are so heavy. But then there's the pandemic quarantine, and I couldn't go back to the dorm and get the books. This is devastating as I need them to finish my dissertation. So, the first thing I did when I got back to school is to digitalize all my materials. I was afraid I wouldn't be able to access them again, so I converted everything to digital format.* (Structure-driven participant P2 in Arabic literature)

Data-driven scholars, on the other hand, usually use born-digital or well-digitised materials, so they don't need to represent the documents while seeking. The data representation is usually embedded in the later data processing activities, which is not in the scope of this paper.

Selecting and verifying

These two characteristics are to ensure the quality, representativeness, and accuracy of the data during focused seeking, and structure-driven scholars' supplementary seeking. In terms of quality, structure-driven participants working with ancient texts are more cautious with the edition selection. They often use the widely acknowledged editions for more trustworthy quality. Participant F1 in art history also mentioned that she would not select images in archaeological reports if they were low-resolution.

Data provenance is important to data-driven scholars using digital sources. They are more confident with sources circulating in their organisations or produced by well-known academic institutes. Participant F5 studying historical geography mentioned that he would only use authoritative databases, such as CHGIS,

> *The project has been going on for many years, I had participated in the database development when I was a student. I'm well informed of the amount of labour, time and money invested in it...A reliable database should be totally transparent with the data processing procedure, standards, and even potential flaws.*

Besides, data-driven scholars also need to select by *representativeness*, as they need to spend considerable time on data processing and analysis, and sometimes it's impossible to process all the available corpora. For instance, participant M1 studying modern Chinese mentioned the need to select high-frequency nouns from existing vocabularies for further analysis. Participant P1 studying ancient Chinese noted,

> *My research focused on three selected classics in Pre-Qin. Whether these three represent Pre-Qin Chinese language is debatable. Some scholars think one text is enough; others believe you need to collect all Pre-Qin literature, which would be impossible to complete even in ten years. So, the compromise is to select some typical works acknowledged by academia, along with some slightly later yet typical works to validate the findings.*

Scholars need to further verify the contents of selected documents or datasets, no matter how much they trust their provenance. Verifying is critical for participants who need to examine exact words and characters, such as scholars of linguistics and philology. Whenever they use digital sources, they are required by academic standards or self-restrained to check back to the physical copies to ensure accuracy and verify page numbers to cite. For example, participant P1 studying ancient Chinese said that he only trusted the textual files circulated within his team, as they had been proofread many times. Even then he would still proofread against the original book by himself to ensure the accuracy. Structure-driven participant D3 also noticed the difference,

> *I wouldn't verify against the original newspaper, because it doesn't matter to me if there's error with one or two words, as long as the meaning is conveyed clearly. But for people studying classics, they wouldn't trust online sources unless they have personally validated and annotated it word by word.*

Monitoring

Unlike monitoring for information across their research phases, only a couple of data-driven scholars concerned with archaeological objects mentioned they would monitor for latest excavations in their supplementary data seeking phase. This is usually conducted at resource level. For example,

> *Many scholars would share newly discovered archaeological items to the Han Painting Database that our team constructed. Though these will not be immediately uploaded to our databases, we have a WeChat subscription account to update such latest news that I would follow for any new images within my data scope.* (Data-driven participant F1 in art history)

In comparison, structure-driven scholars hardly monitor for data. This may be due to their research nature of not emphasising the completeness or extensiveness of data. For example, structure-driven participant D9 who also relies on archaeological objects mentioned,

> *The journals have an annual summary each year, highlighting new discoveries. These are often lengthy articles, and the titles are just annual summaries or yearbooks. It's possible that you spend a lot of time reading, and only to find that they don't contain the data you need. It's a very tedious process, so I wouldn't follow them regularly.*

Discussions and implications

First, the study identified eleven characteristics of humanities scholars' data seeking behaviours, with ten previously recognised in information seeking research and one new characteristic added: representing (*RQ1*). We noticed that elaborations of Ellis's model have introduced extensive components that also covered information managing, processing and use (Makri et al., 2008; Meho & Tibbo, 2003). In this study, we purposefully eliminated these data processing and analysis components due to their complexity in research data contexts. Despite the overlap between data seeking and information seeking characteristics, our study reveals differences that carry implications for data system design and data curation.

We observed that humanities scholars increasingly rely on searching over browsing. As search systems advance, the unique value of browsing now stems less from the ineffectiveness of searching (Buchanan et al., 2005) and more from its ability to offer broader context. Although scholars critically reflect on this point, they still prefer searching when possible. We suggest databases or platforms to support more robust full-text searching capabilities and provide system recommendations based on knowledge inference. This will allow for deeper exploration of documents and compensate for the loss of broader contexts or serendipities in searching. Scholars' emphasis on chaining echoes previous research and highlights the importance of starter references (Duff and Johnson, 2002; Palmer, 2005). Additionally, we found that starter references can also be used to develop data. Thus, their digitisation or tools to facilitate processing them are in great demand. Scholars meet accessing constraints in online databases and libraries and archives. Networking or representing data at the scene are employed to mitigate the constraints. More open resources and encouraging collaboration in institutional or community level would be beneficial. Scholars select resources or documents by data quality and representativeness, and they further verify the contents to ensure accuracy, which was only observed in physicists' information seeking in previous studies (Ellis et al., 1993). Monitoring for data, however, is not as common as information monitoring. This may be due to that many humanities scholars' data are retrospective sources curated as part of humanities research infrastructure (Trace and Karadkar, 2017).

In achieving all the above, the digitisation and curation of the humanities materials are imperative to forge our cultural commonwealth (Poole, 2015), especially the tertiary sources and historical documents such as archives and local gazetteers. Meanwhile, libraries and other curators should be aware of the importance of why and how they select material for digitisation and trace each step of these processes (Oberbichler et al., 2022), as data provenance are particularly important for humanities scholars' data seeking according to our study and previous research (Borgman, 2015).

Second, the study identified two primary research approaches in humanities studies, the data-driven and structure-driven approach. In both approaches, scholars' data seeking consists of three phases – exploratory seeking, focused seeking, and supplementary seeking. The study reveals how data seeking characteristics vary between research approaches and across seeking phases (RQ2).

We mapped humanities scholars' data seeking characteristics to the research approaches and seeking phases. We also highlighted the operating level more prominently (Makri et al., 2008), so that the model is more explicit to derive feasible and actionable design and curation insights. The redefining and restructuring of the seeking characteristics in the perspective of humanities scholars' data seeking marks our primary contribution to the conceptual growth of Ellis's model (Savolainen, 2017).

Awareness of the characteristic differences between research approaches can help tailor data services to target research communities. For example, structure-driven scholars often focus on document-level representation, suggesting that solutions like increasing human resources (Borgman, 2015) or reducing photocopying and scanning costs at libraries and archives could meet their needs until mass materials are digitised. For data-driven scholars who select at resource level, databases or platforms facilitating data-driven research or digital humanities research should clearly document the data provenance, with transparency on data processing procedure, standards, and potential flaws, thereby help scholars make informed resource selections and build trust. At content level, structure-driven scholars mainly take notes to extract data, while some data-driven scholars already leverage computational methods for extraction, such as regular expressions and dictionary-based extraction. These techniques improved efficiency in extracting and can also subsequently transform data into specified structure. Yet many scholars are not equipped with such skills, which urges educations and training in technical skills and data literacy, especially for structure-driven scholars who are considering data-driven approach.

The seeking phases enables a better understanding of the data seeking process (Bronstein, 2007). We found that seeking characteristics in the exploratory phase are quite similar. However, in the more intense focused seeking phase, structure-driven scholars conform to Palmer's (2005) *humanities mode* of information access, which is centrifugal and follows an interpretive course, whereas data-driven scholars' seeking mode is closer to the *scientific model*, which is problem-oriented, directed, and centripetal. While structure-driven scholars continue to seek data actively in the supplementary phase, data-driven scholars mainly monitor for previous missing data. This underscores that data seeking, along with other data interactions, should be examined closely in conjunction with research approaches, methodologies, or processes (Hoekstra and Koolen, 2019). With humanities research increasingly shifting towards data-driven approaches and incorporating computational methods, more attention should be paid to examine humanities scholars' data behaviours.

Conclusion

Based on 27 in-depth interviews with humanities scholars, our study summarised humanities scholars' data seeking characteristics operating at varying levels and identified their variations across different seeking phases and research approaches. The study expanded the utility of Ellis' model beyond its original information seeking contexts and indicated its potential applicability to data seeking. It contributes to the conceptual growth of Ellis' model and provides practical implications in system design and humanities data curation. We note that our participants primarily used textual data, thus the findings may not fully capture seeking behaviours related to various non-textual data in humanities research. This limitation points to future exploration into data behaviours around diverse humanities data types. Furthermore, we expect the study to stimulate broader discussions on data seeking behaviours across disciplines.

Acknowledgements

The authors thank the participants and reviewers. The study was supported partially by NSFC Grant #72174014 and Grant# 72010107003.

About the authors

Wenqi Li is a Ph.D. candidate and an incoming postdoctoral researcher at the Department of Information Management, Peking University. Her research areas include data behaviour, data curation, digital humanities, and user experience. She can be contacted at wenqili@pku.edu.cn.

Pengyi Zhang is an associate professor at the Department of Information Management, Peking University. Her research expertise includes information and knowledge organization, information/data seeking and sensemaking, and user-centred design and evaluation. She is the corresponding author and can be contacted at pengyi@pku.edu.cn.

Jun Wang is a professor at the Department of Information Management, and the director of the Digital Humanities Research Center, Peking University. His research interests include digital humanities, digital libraries, knowledge organization, information behaviours and web product design. He can be contacted at junwang@pku.edu.cn.

References

Al Shboul, M. K., & Abrizah, A. (2016). Modes of information seeking: Developing personas of humanities scholars. *Information Development*, 32(5), 1786–1805. https://doi.org/10.1177/0266666915627673

Anderson, S., Blanke, T., & Dunn, S. (2010). Methodological commons: Arts and humanities e-science fundamentals. *Philosophical Transactions of the Royal Society of London. Series A:*

Mathematical, Physical and Engineering Sciences, 368(1925), 3779–3796. https://doi.org/10.1098/rsta.2010.0156

Borgman, C. L. (2008). Data, disciplines, and scholarly publishing. *Learned Publishing*, 21(1), 29–38. https://doi.org/10.1087/095315108X254476

Borgman, C. L. (2010). *Scholarship in the digital age: Information, infrastructure, and the internet*. MIT Press.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059–1078. https://doi.org/10.1002/asi.22634

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Bronstein, J. (2007). The role of the research phase in information seeking behaviour of Jewish studies scholars: A modification of Ellis's behavioural characteristics. *Information Research*, 12(3).

Brown, C. D. (2002). Straddling the humanities and social sciences: The research process of music scholars. *Library & Information Science Research*, 24(1), 73–94. https://doi.org/10.1016/S0740-8188(01)00105-0

Buchanan, G., Cunningham, S. J., Blandford, A., Rimmer, J., & Warwick, C. (2005). Information seeking by humanities scholars. In A. Rauber, S. Christodoulakis, & A. M. Tjoa (Eds.), *Research and advanced technology for digital libraries* (Vol. 3652, pp. 218–229). Springer Berlin Heidelberg. https://doi.org/10.1007/11551362_20

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2015). Data practices and curation vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology*, 66(3), 616–633. https://doi.org/10.1002/asi.23184

Chu, C. M. (1999). Literary critics at work and their information needs: A research-phases model. *Library & Information Science Research*, 21(2), 247–273. https://doi.org/10.1016/S0740-8188(99)00002-X

Duff, W. M., & Johnson, C. A. (2002). Accidentally found on purpose: Information-seeking behavior of historians in archives. *The Library Quarterly: Information, Community, Policy*, 72(4), 472–496. https://www.jstor.org/stable/40039793

Ellis, D. (1989a). A behavioural appoach to information retrieval system design. *Journal of Documentation*, 45(3), 171–212. https://doi.org/10.1108/eb026843

Ellis, D. (1989b). A behavioural model for information retrieval system design. *Journal of Information Science*, 15(4–5), 237–247. https://doi.org/10.1177/016555158901500406

Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation*, 49(4), 356–369. https://doi.org/10.1108/eb026919

Ellis, D., & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, 53(4), 384–403. https://doi.org/10.1108/EUM0000000007204

Flanders, J., & Muñoz, T. (2012). An introduction to humanities data curation. *DH Curation Guide: A Community Resource Guide to Data Curation in the Digital Humanities*. https://guide.dhcuration.org/contents/intro/

Ge, X. (2010). Information-seeking behavior in the digital age: A multidisciplinary study of academic researchers. *College & Research Libraries*, 71(5), 435–455. https://doi.org/10.5860/crl-34r2

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, 70(5), 419–432. https://doi.org/10.1002/asi.24165

Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice. *Journal of Information Science*, 46(4), 459–475. https://doi.org/10.1177/0165551519837182

Gualandi, B., Pareschi, L., & Peroni, S. (2023). What do we mean by "data"? A proposed classification of data types in the arts and humanities. *Journal of Documentation*, 79(7), 51–71. https://doi.org/10.1108/JD-07-2022-0146

Hoekstra, R., & Koolen, M. (2019). Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 52(2), 79–94. https://doi.org/10.1080/01615440.2018.1484676

Koolen, M., Kumpulainen, S., & Melgar-Estrada, L. (2020). A workflow analysis perspective to scholarly research tasks. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 183–192. https://doi.org/10.1145/3343413.3377969

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5), 361–371. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-%23

Late, E., & Kumpulainen, S. (2022). Interacting with digitised historical newspapers: Understanding the use of digital surrogates as primary sources. *Journal of Documentation*, 78(7), 106–124. https://doi.org/10.1108/JD-04-2021-0078

Li, W., Zhang, P., & Wang, J. (2023). Humanities scholars' understanding of data and the implications for humanities data curation. *Proceedings of the Association for Information Science and Technology*, 60, 1034–1036. https://doi.org/10.1002/pra2.936

Ma, R., & Xiao, F. (2020). Data practices in digital history. *International Journal of Digital Curation*, 15(1), Article 1. https://doi.org/10.2218/ijdc.v15i1.597

Makri, S., Blandford, A., & Cox, A. L. (2008). Investigating the information-seeking behaviour of academic lawyers: From Ellis's model to design. *Information Processing & Management*, 44(2), 613–634. https://doi.org/10.1016/j.ipm.2007.05.001

Meho, L. I., & Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American Society for Information Science and Technology*, 54(6), 570–587. https://doi.org/10.1002/asi.10244

Moulaison-Sandy, H., & Wenzel, A. G. (2023). The records data ecosystem in humanities and human sciences scholarship. *Portal: Libraries and the Academy*, 23(1), 67–88. https://doi.org/10.1353/pla.2023.0000

Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., & Tolonen, M. (2022). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, 73(2), 225–239. https://doi.org/10.1002/asi.24565

Pacheco, A. (2022). Digital humanities or humanities in digital: Revisiting scholarly primitives. *Digital Scholarship in the Humanities*, 37(4), 1128–1140. https://doi.org/10.1093/llc/fqac012

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology*, 56(11), 1140–1153. https://doi.org/10.1002/asi.20204

Palmer, C. L., Teffeau, L. C., & Pirmann, C. M. (2009). *Scholarly information practices in the online environment: Themes from the literature and implications for library service development*. OCLC Research.

Poole, A. (2015). *Forging our cultural commonwealth: The importance of digital curation in the digital humanities*. ProQuest. https://www.proquest.com/openview/4a3c2eeae1882ee5cabf2933ce677aaf/1?pq-origsite=gscholar&cbl=18750

Rhee, H. L. (2012). Modelling historians' information-seeking behaviour with an interdisciplinary and comparative approach. *Information Research*, 17(4). http://informationr.net/ir/17-4/paper544.html#.YMB7AZMza3I

Rolland, B., & Lee, C. P. (2013). Beyond trust and reliability: Reusing data in collaborative cancer epidemiology research. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* - CSCW '13, 435. https://doi.org/10.1145/2441776.2441826

Savolainen, R. (2017). Contributions to conceptual growth: The elaboration of Ellis's model for information-seeking behavior. *Journal of the Association for Information Science and Technology*, 68(3), 594–608. https://doi.org/10.1002/asi.23680

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the humanities. *Journal of Digital Humanities*, 2(3), 2–13.

Schroeder, R. (2014). Big Data: Towards a more scientific social science and humanities? In Mark Graham, and William H. Dutton (eds), *Society and the Internet: How Networks of Information and Communication are Changing Our Lives*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199661992.003.0011

Smith, K. (1988). *An investigation of the information seeking behaviour of academics active in the field of English literature* [PhD Thesis]. University of Sheffield, Department of Information Studies.

Trace, C. B., & Karadkar, U. P. (2017). Information management in the humanities: Scholarly processes, tools, and the construction of personal collections. *Journal of the Association for Information Science and Technology*, 68(2), 491–507. https://doi.org/10.1002/asi.23678

Wang, X., Duan, Q., & Liang, M. (2021). Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology*, 72(9), 1161–1182. https://doi.org/10.1002/asi.24483

Whitmore, D. A. (2016). *Seeking context: Archaeological practices surrounding the reuse of spatial information*. University of California.

Wiberley, S. E., & Jones, W. G. (1989). Patterns of information seeking in the humanities. *College & Research Libraries*, 50(6), 638–645. https://doi.org/10.5860/crl_50_06_638

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation*, 55(3), 249–270. https://doi.org/10.1108/EUM0000000007145

Wilson, T. D. (2000). Human information behavior. *Informing Science*, 3(2), 49–56. https://doi.org/10.28945/576

Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, 68(4), 946–956. https://doi.org/10.1002/asi.23730

Zhang, P., & Soergel, D. (2014). Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. *Journal of the Association for Information Science and Technology*, 65(9), 1733–1756. https://doi.org/10.1002/asi.23125

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, 7(1), 5–16. https://doi.org/10.1007/s00799-007-0015-8

Appendix I: Interview Guide

1. Basic Information

Please introduce your department, major, grade or professional title, and main research focus.

2. Research Interest and Approach

2.1 Please describe the most representative research project you worked on, including the research question, methods or process, team and division of labour, challenges, and value, etc.

2.2 Do you have any understanding or experience with digital humanities research? Please briefly share your thoughts.

3. Research Materials and Data

3.1 What are your research objects and materials used in the above project?

3.2 What is your understanding of "data"? Do the above research objects or materials constitute "data"? Why?

4. Collecting Data

4.1 Please describe the process of collecting research materials in the above study, including sources, methods, difficulties, and whether the results met expectations. Can you share any example photos or screenshots of the above materials?

4.2 There are many online public databases, platforms, and corpora available today. Are you aware of, intend to try, or have experience using any of them? Why?

4.3 Were there any teams or other individuals involved in the collection of materials? What do you think about the collaboration? Please describe the collaborative process and any problems encountered in detail.

5. Using Data

5.1 Please describe how you organize, process, analyse, and write using the research materials in the above study, including methods, technical tools, difficulties, and results. Can you share any example photos or screenshots of the usage process?

5.2 Were there any teams or individuals involved in the aforementioned use of research materials? What do you think about the collaboration? Please describe the collaborative process and any problems encountered in detail.

6. Management and Sharing of Research Materials

6.1 How do you organize and manage research materials in the above study? Can you share screenshots or photos of your data folder or other management methods?

6.2 Do you have a “research material collection” that you maintain and use over the long term? How was it established and maintained?

6.3 Have you ever shared your research materials or personal collection with other scholars, teams, or a broader audience? Why or why not?

6.4 What are your views on the sharing, publishing, and curation of humanities research data? Would you participate in such efforts?

7. Summary

7.1 Please describe your overall process of interacting with the research materials in the above research.

7.2 Do you have any other questions or views you would like to discuss or share with us?

Appendix II: Codebook

Category	Code	Definition	Reference
Concept of Data	Data	Research materials used as the evidence to be processed and analysed to form or validate a theory; objects of interpretation, rebuttal, or scholarly discourse; evidence, example, or quotation to make an argument.	(Li et al., 2023)
	Non-data materials	Research materials used to refer to other sources; to help understand the background or form research questions.	(Li et al., 2023)
Research Approach	Data-driven	In this research approach, scholars discover patterns or form theories inductively based on a substantial amount of data or texts.	(Zhang & Soergel, 2014)
	Structure-driven	In this research approach, scholars first develop a discourse structure based on a certain amount of data, then add data to the structure, and ultimately form observations or interpretations of historical, cultural, or philosophical phenomena.	(Zhang & Soergel, 2014)
Seeking Phases	Exploratory seeking	The seeking and accumulation of data prior to the identification of research questions and information needs.	-
	Focused seeking	The data seeking driven by established information needs after the research questions are identified.	-
	Supplementary seeking	The data seeking that takes place when initial data seeking is complete, and	-

		core ideas are formulated. This phase is usually in parallel with data processing and analysis.	
Seeking Behaviours	Accessing	The process of gaining access to a resource or document.	(Makri et al., 2008)
	Starter reference	Identifying a key document to commence the search.	(Bronstein, 2007)
	Monitoring	Maintaining awareness of an area of interests through regularly following certain sources.	(Ellis & Haugan, 1997)
	Chaining	Following chains of citations or other forms of referential connections between documents.	(Ellis, 1989a)
	Browsing	Semi-directed searching in an area of potential interest.	(Ellis, 1989a)
	Searching	Formulating a query (using keywords, filters, metadata, query builder etc.) to locate data.	(Gregory et al., 2019; Makri et al., 2008)
	Verifying	Checking the information and sources found for accuracy and errors.	(Ellis et al., 1993)
	Selecting	Carefully choosing resources or documents as being potential data sources, usually by assessing quality or provenance.	(Chao et al., 2015; Makri et al., 2008)
	Extracting	Systematically working through a resource or document to identify potential data.	(Makri et al., 2008)
	Networking	Obtain data from colleagues by collaboration or informal personal requests.	(Bronstein, 2007; Chao et al., 2015; Gregory et al., 2019)
	Representing	Representing research materials in means of taking notes, taking photographs, scanning, etc.	-
Level	Resource	A resource contains many documents that can be accessed and further used as data, such as a library, archive, database, and repository.	(Makri et al., 2008)
	Document / Dataset	A document or dataset is the unit of access and can be used (e.g.: extracted, processed, or analysed) as data.	(Makri et al., 2008)
	Content	The actual content or data within the documents or datasets. This involves in-depth reading, selecting and extraction of specific data, or insights from the documents.	(Makri et al., 2008)