

# Information-Seeking in Large-Scale Digital Libraries

## Strategies for Scholarly Workset Creation

David M. Weigl, Kevin R. Page

Oxford e-Research Centre

University of Oxford

United Kingdom

{david.weigl,kevin.page}@oerc.ox.ac.uk

Peter Organisciak, J. Stephen Downie

School of Information Sciences

University of Illinois at Urbana-Champaign

United States

{organis2,jdownie}@illinois.edu

### ABSTRACT

Large-scale digital libraries such as the HathiTrust contain massive quantities of content combined from heterogeneous collections, with consequential challenges in providing mechanisms for discovery, unified access, and analysis. The HathiTrust Research Center has proposed ‘worksets’ as a solution for users to conduct their research into the 15 million volumes of HathiTrust content; however existing models of users’ information-seeking behaviour, which might otherwise inform workset development, were established before digital library resources existed at such a scale.

We examine whether these information-seeking models can sufficiently articulate the emergent user activities of scholarly investigation as perceived during the creation of worksets. We demonstrate that a combination of established models by Bates, Ellis, and Wilson can accommodate many aspects of information seeking in large-scale digital libraries at a broad, conceptual, level. We go on to identify the supplemental information-seeking strategies necessary to specifically describe several workset creation exemplars.

Finally, we propose complementary additions to the existing models: we classify strategies as instances of *querying*, *browsing*, and *contribution*. Similarly we introduce a notion of *scope* according to the interaction of a strategy with *content*, *content-derived metadata*, or *contextual metadata*. Considering the scope and modality of new and existing strategies within the composite model allows us to better express—and so aid our understanding of—information-seeking behaviour within large-scale digital libraries.

### CCS CONCEPTS

•Information systems →Users and interactive retrieval; Retrieval tasks and goals;

### KEYWORDS

information-seeking behaviour, digital libraries, workset creation

## 1 INTRODUCTION

The HathiTrust Digital Library (HTDL) provides access to over 14.8 million digitized volumes, comprising approximately 5.2 billion

pages. The sheer scale of this resource, operating over heterogeneous content while providing combined access to a diverse set of collections, entails new challenges in supporting discovery, analysis, and scholarly use of the wealth of materials contained within.

A study of how potential HTDL users maintain collections for research [9] identified functionalities required by academics when working with large-scale digital libraries. **Worksets** seek to address these needs, offering capabilities such as: gathering individual documents within citable collections; supporting computational extraction of document-derived features for scholarly analysis; generating new descriptive information by inference over existing metadata; and linking complementary information from external sources. Jett [11] implements worksets as machine-actionable research collections aggregating members, metadata intrinsic to the workset’s essential nature, metadata intrinsic to digital architectures, metadata supportive of human interactions, derivative metadata from workset members, and metadata concerning workset provenance.

The explicit intent of worksets is to support scholarly research; thus, ongoing development is strongly influenced by the available literature on scholars’ information-seeking behaviours. However existing models of information-seeking behaviour were established well before digital libraries at the scale of HTDL. In this paper we assess the coverage of these models when applied to large-scale digital libraries, and strategies which might enable them to accommodate the affordances of systems such as the HTDL.

## 2 BACKGROUND

Information-seeking behaviour is a well-studied field of information science [15]. Many models and conceptual frameworks have been proposed, partly due to a tendency of model creation over elaboration [14]. Most well-established models were largely developed in the 1980s-90s [3]; we discuss the most pertinent here.

Ellis’s highly influential model of information-seeking behaviour in the research process [6][8] was derived from interviews with social scientists, physicists, chemists, and English literature researchers. It maps behaviours to eight **categories**: *starting* characteristics of the initial search, e.g. references, authors, or broad queries on search tools; *chaining*, traversal of references and citations; *browsing*, semi-directed or -structured searching of an area of potential interest; *differentiating* between sources, e.g. by approach, author, or publication venue; *monitoring* of a (sub-)field to keep one’s knowledge up to date; *extracting* selected material from a source for closer consideration; *verifying* the accuracy of information obtained, e.g. by reference to external sources; and *ending*, information gathering toward the end of a project, e.g. to situate it within the wider literature.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

JCDL2017, Toronto, ON, Canada

© 2017 Copyright held by the owner/author(s). 978-x-xxxx-xxxx-x/YY/MM...\$15.00  
DOI: 10.1145/nnnnnnn.nnnnnnn

Informed by Ellis, Bates’s “berrypicking” model [12] casts information seeking as an evolving process of iterative refinement and reconception, where each encountered piece of information provides new ideas or directions. The user chases leads between different sources using a range of **strategies**. Bates describes six, noting the set is not exhaustive: *subject search*; *author search*; *journal run*, inspection of all articles published in a journal for a given timespan; *area scanning*, browsing of material collocated with previously identified items (cf. physical library shelf); *backward chaining* by following references; and *forward chaining* by finding material citing a given item. The user’s information need is not satisfied by a single set of results, but rather by a series of selections of references and bits of information – analogous to berries, scattered among bushes in the forest, being picked one handful at a time.

Different conceptions of **modes** have been established to further account for the motivations and contexts of information seeking. Choo, Detlor, and Turnbull [4] describe four modes of browsing and searching, identified in work on environmental scanning [1][13]: *formal search* guided by a specific information need; *informal search*, unstructured searching to deepen one’s understanding of a specific issue; *conditioned viewing* of information about selected topics in order to evaluate significance; and *undirected viewing*, broad scanning with no specific information need in mind. Wilson [15] offers an alternative set of four modes: *active search*, actively seeking out information; *passive search*, where information-seeking behaviour results in the incidental acquisition of relevant information that was not actively sought; *ongoing search* to update or expand one’s established knowledge; and *passive attention*, wherein knowledge acquisition takes place coincidentally, outside of any information-seeking context. Notably, *passive attention* explicitly accounts for serendipitous discovery, whereas the loosely equivalent *undirected viewing* of the other model demands a “considerable degree of orientation” in the selection of particular sources [1].

The above models were established before the widespread availability of massive-scale distributed information systems. However, they can still support how we look at modern systems, and inspire novel solutions to information-seeking problems. In the early days of the World Wide Web, Choo et al [4] cross-tabulated Ellis’s categories with the four modes of environmental scanning to establish a model of the behaviours of information professionals using this novel resource. They describe a set of “Web Moves” afforded by this environment: identifying web sites; following links; scanning site maps; means of selection (e.g. bookmarking, printing); receiving push updates; searching within pages; and using search engines.

The process of building worksets is particularly underserved by a view of information-seeking behavior in terms of single-query sessions. Although the strategies offered by Bates as exemplars of berrypicking do not encompass capabilities of the HTDL and other large-scale digital libraries – e.g. feature extraction and use, metadata enrichment – workset uses are analogous to the behaviour characterised by “berrypicking”. Rather than developing new models, we recontextualize and elaborate on the existing models in the context of workset building strategies.

In the remainder of this paper, we consider the implications of established and novel strategies for exploration, discovery, and scholarly analysis of the HTDL’s diverse collections and datasets.

**Table 1: Overlap of information-seeking models**

Category \ Strategy	Starting	Chaining	Browsing	Differentiating	Monitoring	Extracting	Verifying	Ending
Subject search	A		A/P	*	O			A/P
Author search	A		A/P	*	O			A/P
Journal run	A		A/P		O	A/P		A/P
Area scan			A/P		O			
Backwards chaining		A/P	A/P	*			A/P	
Forwards chaining		A/P	A/P	*	O			

Modes of searching—A: Active search, P: Passive search, O: Ongoing search.

\*: May be used to differentiate when filtering (but not as a search strategy)

### 3 INFORMATION SEEKING IN THE HTDL

#### 3.1 Strategies for scholarly research

We investigate the coverage of Bates’s strategies within a framework of information-seeking behaviour offered by the models introduced in section 2, considered in terms of their application to the large-scale resource for scholarly research offered by the HTDL.

The reviewed literature supports a strong relationship between categories of behaviour within the research process (acc. Ellis) and the features of information seeking employed. To explore this relationship further, we adapt Choo et al’s approach of cross-tabulation to investigate correspondences between Ellis’s categories and the strategies described by Bates. Like Choo, we are interested in understanding modes of information seeking available through these interactions, but rather than applying the modes of environmental scanning, we apply those of Wilson [15] so to incorporate notions of serendipitous discovery (*passive attention*).

The results of our cross-tabulation of Bates’s strategies and Ellis’s categories, with Wilson’s modes of searching, are displayed in Table 1. As expected, there is a differentiated correspondence of applicable strategies and modes of searching across the categories of the information-seeking process. Ellis’s *starting* category presupposes an actively pursued information need, although it may be loosely defined; hence, facet-based explorations according to subject, author, or publication venue are possibilities; whereas the *area scan*, and *backward* and *forward chaining* strategies require a starting point to already have been established in earlier search. The correspondence of the *chaining* category and strategies is straightforward, affording opportunities for active and passive search. Each strategy affords opportunities for *browsing*: sought information may be identified, and incidentally useful information encountered, by browsing through records along facet-based navigation vectors; by *area scanning*, either physically along a library shelf, or digitally by virtue of analogous catalogue metadata (e.g. call number); or via contextual traversal of a citation chain. *Monitoring* corresponds straightforwardly to *ongoing search*. Ellis’s respondents suggest that this is predominantly achieved by following trusted publications, making the *journal run* the most applicable strategy. However, in principle any search strategy can be iteratively reapplied with a view toward acquiring newly available information, with the exception of *backward chaining* (as articles do not accumulate further references once published). *Extracting* is supported by the *journal run* strategy, which entails purposeful selection of documents from a larger stream of potentially relevant material. *Backward chaining* may be applied to follow up claims

in referenced articles (*verifying*). Finally, activities in the *ending* category are defined largely in relation to *starting* [7], although the supplementary nature of research at this stage offers greater opportunity for discoveries of incidentally useful information (*passive search*).

Several gaps between the models' coverage become apparent. None of Bates's strategies address *differentiating*; *extracting* and *verifying* are limited. While Bates acknowledges her strategies are not intended to be comprehensive, these gaps also correspond to less comprehensive support for differentiation, extracting, and verifying within information systems available in 1989. Further, Wilson's *passive attention* does not apply to any of the investigated correspondences, as each presupposes the initial presence of an information need. Yet, differentiation and verification are especially important in collections the size of the HathiTrust, where techniques for data reduction, disambiguation, and duplicate detection assist in promoting precision when recall is very high; extraction, afforded by the ability to reference information at large (e.g. workset-level) and fine (e.g. paragraph-level) grains of granularity, supports iterative scholarly research processes; and serendipitous discovery is a key affordance of the increasing quantity and ubiquity of information afforded by modern technologies [5].

### 3.2 Strategies for workset creation

To elicit the deficiencies of coverage identified in the preceding section, we now consider illustrations of research using HTDL worksets. We identify enacted information-seeking strategies *in italics*, including applications of Bates's strategies; and of additional strategies not covered by Bates *in bold italics*.

**W1: Novels by Austen's contemporaries.** An English literature scholar conducts an *author search* to identify Austen's works; then applies *contextual browsing*, a form of *chaining* operating over any available metadata, refining to contemporaneous works according to the publication dates of Austen's. Finally, she specifies a "language and literature" classification (*subject search*), yielding a workset that can be refined, analysed, and referenced in publication.

**W2: Perception of labour unions in different countries.** A social historian prepares a manuscript on differing cultural attitudes towards labour unions. She conducts a *concept search* [10] on the term "union". Contextual metadata maps the word to a disambiguated 'concept' encompassing "labour organisation" while excluding e.g. "political union" or "union" in mathematical set theory. She then *browses by publication place* in the resulting collection.

**W3: Bicycle illustrations within the HathiTrust.** A computer vision algorithm is applied on a workset to identify pages containing bicycles illustrations (*feature generation*). Image coordinates are stored as contextual metadata (*entity annotation*).

**W4: Bicycle illustrations in 19th century British works.** A curator identifies British works published in the 19th century (*publication date* and *publication place search*). She then performs a *feature-based search* on contextual metadata of the type generated in W3 to extract instances of bicycle illustrations within this workset. The resulting images are inspected (*content browsing*) to select particularly illustrative examples for use in an exhibit.

**W5: Collecting dictionary duplicates to evaluate OCR.** A computational linguistics researcher seeks scans of duplicate resources to evaluate the performance of optical character recognition

(OCR) algorithms. He performs a *genre search* on "encyclopaedias and dictionaries", narrowing down to instances of "Webster's" via a *title search*. Browsing according to publication details (*browse by publication place* and *date*), he identifies and marks a *duplicate annotation* over several scans of "Websters new international dictionary of the English language" published in 1952.

#### W6: Studying cultural skews in attitudes towards London.

A sociologist performs a *concept search* to disambiguate the term 'London', removing irrelevancies (e.g. 'London, Ohio'). She uses previously generated feature information to remove results situated on paratextual pages. Finally, she *browses by publication place*, investigating cultural skews in attitudes towards the British capital.

**W7: The notion of relevance in Information Science.** A doctoral student wishes to understand influential conceptualizations of relevance. Performing a *subject search* on 'Library and Information Science', followed by a *keyword search* for "relevance", he ranks the results by number of citations. He then contextually browses the articles citing, or referenced in, the most highly cited works, iterating amongst the retrieved documents.

## 4 A MODEL OF INFORMATION SEEKING FOR LARGE-SCALE DIGITAL LIBRARIES

Tabulating the emergent information-seeking strategies (section 3.2) within our framework (section 3.1) in Table 2 we now consider information seeking for workset creation within the HTDL.

Bates's strategies are expressed granularly, e.g. *author search* over the more general *catalogue search*. The increased number of strategies identified in section 3.2 corresponds to increased technical capabilities of large scale DLs. We generalise these strategies into three behaviours classes: *querying*, *browsing* and *contributing*, the former two being well-described in the literature.

*Contributing* by user (e.g. annotations, workset assignment) or system (e.g. extracted features, provenance traces) is a new addition required to articulate iterative use and reference in workset creation and analysis. As a relatively recent technological affordance, *contributing* is insufficiently described in previous information-seeking models. We further distinguish different *scopes* over which information-seeking strategies are applied. These include *content* (scanned images, textual content obtained via OCR); *content-derived metadata* determined from the content alone; and *contextual metadata*, supplemented by system or user processes<sup>1</sup>.

Having established these behavioural classes and scopes of application, we observe patterns in their interaction: *querying* and *browsing* strategies apply across all scopes; whereas *contributing* is limited to *contextual metadata* (by definition – all contributions must involve generative user and/or system actions). For our examples, Choo et al's strategies provide adequate coverage of the *content* scope; and Bates's partially for *content-derived*. We note every *new* information-seeking strategy is applied within *content-derived* and *contextual metadata* scopes, which are particularly relevant to dynamic collections. By implication, our elaborations of existing models are required to fully articulate and understand information-seeking behaviour in the context of contemporary large-scale digital libraries.

<sup>1</sup>The *content* scope is scarcely represented within Table 2 as an effect of brevity, as most worksets ultimately result in content inspection by the user.

**Table 2: Information-seeking strategies for workset creation**

Category		Starting	Chaining	Browsing	Differentiating	Monitoring	Extracting	Verifying	Ending	Scope of application	W1	W2	W3	W4	W5	W6	W7
Query	Strategy																
	Keyword search	A		A/P	*	O			A/P	Content							•
	Author search	A		A/P	*	O			A/P	Content-Derived	•						
	Title search	A		A/P	*					Content-Derived					•		
	Publication place	A		A/P	*	O			A/P	Content-Derived				•			
	Publication date	A		A/P	*	O			A/P	Content-Derived				•			
	Genre search	A		A/P	*	O			A/P	Content-Derived					•		
	Subject search	A		A/P	*	O			A/P	Content-Derived	•						•
	Concept search	A		A/P	A	O			A/P	Contextual		•				•	
Browse	Feature-based search	A		A/P	*	O			A/P	Contextual				•		•	
	Content browsing			A/P	A		A	A/P		Content				•			
	Publication place	A	A/P	A/P	A		A		A/P	Content-Derived	•				•	•	
	Publication date	A	A/P	A/P	A		A		A/P	Content-Derived					•		
Contribute	Contextual browsing	S	A/P/S	A/P	A		A	A/P	A/P	Contextual	•						•
	Feature generation	A			A	O	A			Contextual			•				
	Entity annotation				A			A		Contextual			•				
	Duplicate annotation				A			A		Contextual					•		
	Workset instantiation		A/P	A/P	A	A/P	A	A	A	Contextual	•	•	•	•	•	•	•

Modes of search—A: Active search, P: Passive search (P), O: Ongoing search, S: Serendipitous discovery (*passive attention*), \*: Differentiates when filtering

#### 4.1 Summary of the combined model

We propose a model of information-seeking behaviour for scholarly workset creation in large-scale digital libraries, combining Bates’s information-seeking **strategies** [12], **categories** of information-seeking behaviour within the research process per Ellis [7], and Wilson’s **modes** of information seeking [15].

In addition, we classify strategies as **querying**, **browsing**, and **contributing** behaviours. The strategies operate upon **scopes** of application:

- content** representations (e.g. images, OCR’d text),
- content-derived metadata**, derived from content alone,
- contextual metadata**, derived via external processes.

#### 5 FUTURE WORK AND CONCLUSIONS

In future work we will apply our model to differentiate user affordances offered by candidate technologies for implementing large-scale digital libraries, such as enterprise search databases and semantic web triplestores, and validate the model through user trials.

We will also extend the conceptualization presented here in relation to agency and serendipitous discovery (Wilson’s *passive attention*). The strategies presented in this paper are initiated by a human information seeker; however, we posit there may be parallel, system-initiated information filtering [2] actions, either requested by the user as a profile-derived “deferred query”, or automatically through algorithmic heuristics. Consequently, under agency of the system, any of the presented strategies may serve the task of *differentiating* (Table 2), a core purpose of information filtering.

Serendipitous discovery plays a role in only one of the presented strategies (*contextual browsing*). This is not to diminish the importance of the concept; rather, the focus on *strategies* precludes further attention, as serendipitous discovery is—almost by definition—not a strategy. But it is clearly an important affordance of ubiquitous digital information [5]. Furthermore, when *passive attention* results in the serendipitous acquisition of relevant information, it may in turn trigger *active searching* activity.

We have presented a model of information-seeking behaviour focussing on strategies for scholarly workset creation in the HTDL.

It combines aspects of three well-established models in the literature, introduces additional *contributing* strategies for information seeking, and specifies three scopes over which information-seeking strategies may be applied. These elaborations support the articulation of information-seeking behaviours concomitant to the emergent properties of large-scale digital libraries.

#### ACKNOWLEDGMENTS

Undertaken through the Workset Creation for Scholarly Analysis and Data Capsules project, funder: Andrew W. Mellon Foundation.

#### REFERENCES

- [1] F. J. Aguilar. 1967. *Scanning the business environment*. Macmillan.
- [2] N. J. Belkin and W. B. Croft. 1992. Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM* 35, 12 (1992), 29–38.
- [3] Donald O. Case. 2012. *Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior*. Emerald Group Publishing.
- [4] C. W. Choo, B. Detlor, and D. Turnbull. 2000. Information seeking on the Web: An integrated model of browsing and searching. *First Monday* 5, 2 (2000).
- [5] O. de Bruijn and R. Spence. 2001. Serendipity within a ubiquitous computing environment: A case for opportunistic browsing. In *UbiComp 2001*. 362–369.
- [6] D. Ellis. 1989. A behavioural approach to information retrieval system design. *Journal of documentation* 45, 3 (1989), 171–212.
- [7] D. Ellis. 1993. Modeling the information-seeking patterns of academic researchers: A grounded theory approach. *The Library Quarterly* (1993), 469–486.
- [8] D. Ellis, D. Cox, and K. Hall. 1993. A comparison of the information seeking patterns of researchers in the physical and social sciences. *Documentation* 49, 4 (1993), 356–369.
- [9] K. Fenlon, M. Senseney, and et al. 2014. Scholar-built collections: A study of user requirements for research in large-scale digital libraries. In *In Proc. ASIST*.
- [10] A. Hinze, C. Taube-Schock, D. Bainbridge, R. Matamua, and J. S. Downie. 2015. Improving Access to Large-scale Digital Libraries Through Semantic-enhanced Search and Disambiguation. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM Press, 147–156.
- [11] J. Jett, T. W. Cole, C. Maden, and J. S. Downie. 2016. The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data* 2 (2016).
- [12] M. J. Bates. 1989. The Design of Browsing and Berrypicking Techniques. *Online review* 13, 5 (1989), 407–424.
- [13] R. L. Daft and K. E. Weick. 1984. Toward a Model of Organizations as Interpretation Systems. *Academy of Management Review* 9, 2 (1984), 284–295.
- [14] R. Savolainen. 2016. Contributions to conceptual growth: The elaboration of Ellis’s model for information-seeking behavior. *Journal of the Association for Information Science and Technology* (2016).
- [15] T. D. Wilson. 1997. Information behaviour: An interdisciplinary perspective. *Information Processing & Management* 33, 4 (July 1997), 551–572.