OXFORD

*Interacting with Computers*, 2025, 1–22
https://doi.org/10.1093/iwc/iwaf001
**Article**

# Introducing the INSPIRE Framework: Guidelines From Expert Librarians for Search and Selection in HCI Literature

Joseph Tu[1,*], Lennart Nacke[1] and Katja Rogers[2]

[1]Stratford School of Interaction Design and Business, System Design Engineering, University of Waterloo, 200 University Ave W, M5H 3C6, Ontario, Canada
[2]Digital Interactions Lab, Informatics Institute, Faculty of Science, University of Amsterdam, Science Park 900, 1098 XH Amsterdam, Netherlands
*Corresponding author: joseph.tu@uwaterloo.ca*

Formalized literature reviews are crucial in human–computer interaction (HCI) because they synthesize research and identify unsolved problems. However, current practices lack transparency when reporting details of a literature search. This restricts replicability. This paper introduces the INSPIRE framework for HCI research. It focuses on the search stage in literature reviews to support a search that prioritizes transparency and quality-of-fit to a research question. It was developed based on guiding principles for successful searches and precautions advised by librarian experts in HCI (n=8) for search strategies in (primarily systematic) literature reviews. We discuss how their advice aligns with the HCI field and their concerns about computational AI tools assisting or automating these reviews. Based on their advice, the framework outlines pivotal stages in conducting a literature search. These essential stages are: (1) defining research goals, (2) navigating relevant databases and (3) using searching techniques (like divergent and convergent searching) to identify a set of relevant studies. The framework also emphasizes the importance of team involvement, transparent reporting, and a flexible, iterative approach to refining the search terms.

**RESEARCH HIGHLIGHTS**

- Introduces the INSPIRE framework to enhance transparency and quality-of-fit during the search stage of literature reviews in HCI.
- Developed the framework based on advice from librarian experts (n=8) and guiding principles for effective search strategies.
- Applies the double-diamond approach, emphasizing divergent and convergent searching techniques to identify a comprehensive and relevant set of studies.

*Keywords*: *research synthesis; literature review; systematic review; artificial intelligence; AI assisted tool; AI screening; manual screening.*

## 1 Introduction

One of the key challenges in conducting SLRs within HCI is the complexity of the search and selection stages. Unlike more homogeneous fields such as medicine, where structured methodologies for SLRs are well established, HCI's interdisciplinary nature makes it difficult to develop standardized approaches. The sheer diversity of topics, methodologies and terminologies can lead to inconsistent search results, making it harder to ensure thorough coverage of relevant literature (Romanelli *et al.*, 2021). Unlike the homogeneous medical field, HCI's heterogeneity complicates search and selection. Although SLRs often provide the final search string used, they neglect to document the process of how they created the search terminology (keystring) and selected the final corpus (Bannigan & Watson, 2009, Chetwynd, 2022). This lack of transparency constrains the replicability of research in the search and selection stages of literature reviews. Thus, we need clear methods to direct our search and provide sound advice to guide researchers through a systematic and replicable search process for HCI research papers.

For a long time, researchers have explored methods for computationally supporting or augmenting the time-consuming literature review process. This research interest is now driven at a faster pace, owing to the increased interest in AI technologies. Although some tools *claim* to have a fully automatic process, studies have shown that human validation is required (Felizardo & Carver, 2020, Monarch, 2021, O'Mara-Eves *et al.*, 2015, Van Dinter *et al.*, 2021a). As Marshall & Wallace (2019) stated, for review stages that involve degrees of subjectivity (e.g. critically assessing a paper's reported study), practitioners may prefer extensive involvement of or full control via an expert human over a machine. For instance, the search and selection stages require careful consideration of the choice of key papers for pre-testing a search strategy, in the critical assessment of screening criteria for a specific research question and in the nuanced assessment of the fit of specific papers to defined selection criteria (described in Section 5). Although computational tools can assist in these tasks, their effectiveness and limitations within the diverse HCI domain remain unclear. An initial framework is needed to guide researchers in a systematic and replicable manner to selecting HCI research papers; balancing human expertise with computational capabilities. Understanding where, when and to what extent human expertise and subjective appraisal should be

**Received:** June 28, 2024. **Revised:** November 25, 2024. **Accepted:** January 12, 2025

valued over computational tools in this process is a challenging open question that reflects the pervasive predicament pulsing at the heart of HCI.

As a first step towards supporting researchers in these stages of search and selection, we draw on established HCI methods to gather and formalize the knowledge of experts through an in-depth interview study with ($n = 8$) librarians at academic institutions with specialized expertise in conducting and guiding formalized literature reviews.

Based on this work, we contribute the initial **INSPIRE framework**, designed to supplement methodological best practices for the search stage in formalized literature reviews (including SLRs) within HCI. Thus, the framework is grounded in librarians' specialized knowledge and skills, including information retrieval, database searching and literature organization expertise (Meert *et al.*, 2016, Spencer & Eldredge, 2018).

The research questions were as follows:

- **RQ1a:** What are experts' recommended best practices for the search and selection stages in formalized literature reviews to aim for rigorous results?
- **RQ1b:** What are experts' perspectives regarding the role of computational tools in the search and selection stages of formalized literature reviews?

Our study contributes to the field of human–computer interaction (HCI) by developing and presenting the initial INSPIRE framework, a structured approach specifically for the search in HCI-based formalized literature reviews. This framework addresses the critical stages of search and selection in a formalized literature review process. These areas have been identified as particularly challenging in existing methodologies (Rogers *et al.*, 2024) and outline key segments[1] for conducting search and selection. These segments include: defining research goals, navigating relevant databases and using complementary search techniques, such as divergent and convergent searching, to ensure a set of relevant studies is identified. Additionally, the framework emphasizes the importance of team involvement, transparent reporting and the regular revisiting of search terms to maintain a thorough scope throughout the process.

## 2 Related Work

Research synthesis is a major challenge in scholarly work and one in which a more formalized methodological revolution is overdue in our field of human–computer interaction (HCI) (Rogers & Seaborn, 2023). Formalized literature reviews[2] hold an important role across various research domains, including HCI, to provide synopses of the current state of the literature but also to help identify gaps, trends, challenges and emerging themes, as well as new knowledge that can inform the development of future technologies and interaction forms (Rogers & Seaborn, 2023, Stefanidi *et al.*, 2023).

For example, a systematic literature review (SLR) can summarize existing scholarly works and publications relevant to a specific topic to provide a more accurate, robust answer to a targeted research question using all available evidence (Uman, 2011). Work done by Stefanidi *et al.* (2023) suggested that many SLRs refer to reporting standards like Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol[3] (Page *et al.*, 2021a, Page, 2021b) or Quality of Reporting of Meta-analyses (QUOROM)[4] (Moher *et al.*, 1999) or provide a flowchart to depict the selection process. However, by virtue of these traits, SLRs in HCI would benefit from a detailed analysis of reporting quality and may need to be adapted to their own reporting guidelines (Rogers *et al.*, 2024).

Traditional methods involve manual searches and selection processes. However, the ever-growing volume of research publications has prompted the exploration of alternative automated approaches (Chai *et al.*, 2021, Reis *et al.*, 2023, Van Dinter *et al.*, 2021a, b, Zhang *et al.*, 2022). For example, Ros *et al.* (2017) addressed the challenges of manual search and selection in SLRs by proposing a semiautomated approach using machine learning. Their method involves training a classifier on an initial set of relevant studies, and then using that classifier to search for and select additional studies iteratively. A researcher validates the top selections using the classifier, and the process continues until a stopping point is reached. This approach has the potential to significantly reduce the manual workload associated with SLR searches, while also improving replicability through a more standardized process. A similar process is followed by ASReview (Van den Brand & van de Schoot), which reorders the stack of corpus papers to be screened based on relevancy, which is informed by a set of human-labelled papers. Manual screening can then be applied; after a specific (pre-defined) number of papers are considered irrelevant, the researchers can then ignore the remaining papers in the stack without viewing them.

Other approaches focus on more specific aspects of searching for papers, such as reference chaining. Chapman *et al.* (2010) introduced a new method to efficiently gather relevant citations by utilizing Scopus' reference list export feature. This approach saves time by automatically extracting citations from articles identified for manual search, eliminating the need for manual identification of new references.

Despite these and similar solutions, the unique characteristics of HCI research pose challenges for traditional review methods. HCI research draws from a wide range of disciplines, including psychology, computer science, design and social sciences (Rogers *et al.*, 2024). However, there is a plethora of research employed, where some studies use surveys, others conduct interviews and so on, and thus creating one search strategy to capture the relevant articles becomes extremely difficult.

Unlike medical research, where randomized clinical trials are the norm or at least much more common (Friedman *et al.*, 2015), HCI research includes unique design applications and methods or study designs informed by a variety of fields, making the replication of results (and indeed the definition of what constitutes a result) particularly challenging (Echtler & Häußler, 2018, Feger *et al.*, 2019, Hornbæk *et al.*, 2014). This contributes to the broader scientific challenge known as the '***reproducibility crisis***' the inability to replicate the results of experiments. Further, the field moves so quickly that even a well-conducted review can quickly become outdated (Elliott *et al.*, 2014, Shojania *et al.*, 2007). This is particularly problematic because systematic reviews are intended to be a comprehensive and reliable source of evidence to inform decision-making (e.g. for the design of future technology).

[1] These segments are not intended to follow a rigid, sequential order; rather, they represent an iterative and often parallel process, where stages may overlap, pause or require revisiting.

[2] With this, we distinguish standalone pieces of research synthesis from the kinds of literature reviews one finds in the related work or background section of research papers that primarily present other contributions.

[3] PRISMA provides a checklist and a flow diagram that authors can use to ensure they include all the essential elements of a systematic review, such as how studies were identified, selected and synthesized.

[4] QUOROM is a set of guidelines for improving the quality and transparency of reporting in meta-analyses.

In fields like medicine, a literature review that is out-of-date risks overlooking critical new findings (Shojania *et al.*, 2007).

## 2.1 Automated Approaches With Computational Tools

Recent advances in AI have led to the development of tools that can assist with various stages of the systematic review process (Blaizot *et al.*, 2022, Khraisha *et al.*, 2024, Mahmoudi *et al.*, 2024, Van Dijk *et al.*, 2023). These tools can be helpful for tasks such as developing search strategies, locating relevant articles, screening and extracting data and even drafting plain-language summaries. However, it is not clear what the tools classify as 'relevant searching' when locating articles (Khalil *et al.*, 2022). Researchers generally agree that AI tools offer significant potential to improve the efficiency and effectiveness of systematic reviews (Allot *et al.*, 2021, Alshami *et al.*, 2023, Brown *et al.*, 2014, Gates *et al.*, 2018, Marshall & Wallace, 2019, Przybyła *et al.*, 2018). However, it is important to acknowledge the potential limitations of these tools when applied at various stages of a formalized literature review (Beller, 2018, Khalil *et al.*, 2022, Tsafnat *et al.*, 2014). For example, Alshami *et al.* (2023) investigated ChatGPT's capabilities across several tasks, including search strategy development, article screening, information extraction and content analysis. Their work suggests that AI can potentially expedite these activities, leading to faster review completion and quicker access to synthesized evidence, in which they achieved 88% accuracy for article screening compared to expert assessment—how much accuracy we should require as a field, however, is unclear. Work by Khalil *et al.* (2022) identified promising tools like LitSuggest (Allot *et al.*, 2021) and Rayyan (Ouzzani *et al.*, 2016) for automating systematic reviews, offering user-friendly support for various review stages, though it should be noted that this was based on self-reporting by authors of review papers rather than systematic comparisons of how well these tools work compared to traditional or manual review methods (Valizadeh *et al.*, 2022).

Limitations exist at various stages and likely to different degrees (Rogers *et al.*, 2024), and may be particularly pronounced with newer tools that still lack established validation and risk assessment (Armijo-Olivo *et al.*, 2020, Cleo *et al.*, 2019, Marshall & Wallace, 2019). In addition, database limitations further restrict access to the full range of relevant sources for specific fields like HCI research, particularly when HCI-relevant work may be scattered across many different disciplines and databases (Allot *et al.*, 2021, Choong *et al.*, 2014). Perhaps most concerning is the bias inherent in some AI tools, often geared towards medical research, which raises concerns about overlooking relevant HCI studies (Brown *et al.*, 2014, Cierco Jimenez *et al.*, 2022, Gates *et al.*, 2018, Marshall & Wallace, 2019, Przybyła *et al.*, 2018). This focus on health-related fields limits generalizability, making AI recommendations less applicable to HCI research questions and design-related interactions (Chai *et al.*, 2021, Clark *et al.*, 2020, Gartlehner *et al.*, 2019, Gates *et al.*, 2020, Pham, 2021).

## 2.2 Guidelines, Protocols, Handbooks for Searching

Many HCI SLRs refer to protocols developed in the medical context for reporting SLRs, most commonly the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol (Page *et al.*, 2021a, Page, 2021b), while PRISMA provides a valuable framework for transparent reporting, its primary focus on intervention-focused studies may limit its applicability to the unique characteristics of HCI SLRs (Stefanidi *et al.*, 2023), though its checklist and flow diagram can still offer a useful foundation for structuring and reporting HCI reviews by outlining essential elements such as study identification, selection and synthesis. However, also guidance specifically for search-related aspects of SLRs. For example, the PRISMA-S extension to PRISMA, which provides a checklist for reporting the search that was conducted (Rethlefsen, 2021), though in our experience, this is only rarely referenced in HCI literature.

Similarly rarely seen in HCI, there are protocols that provide structure not only for reporting but also for the search process itself, such as the Population, Intervention, Comparator, Outcome (PICO) method[5] (Schardt *et al.*, 2007). This contrasts with other approaches like the National Institute for Health and Care Excellence (NICE)[6] (Martinez Garcia *et al.*, 2014); Sample, Phenomenon of Interest, Design, Evaluation, Research type (SPIDER)[7] (Cooke *et al.*, 2012); the Joanna Briggs Institute (JBI)[8] (Jordan *et al.*, 2019, Munn, 2019); and Peer Review of Electronic Search Strategies (PRESS)[9] (McGowan *et al.*, 2016), which provide guidance on the selection of key search terms, synonyms, boolean operators and the combination of search terms. However, none of these guidelines explicitly illustrate the process of '***creating the search keywords for the keystring***'. Instead, they often recommend that the primary search be conducted by a librarian and peer-reviewed by another person (McGowan *et al.*, 2016), which is, to our knowledge, very rare in HCI. To bridge the current gap in the literature on automated search strategies, understanding how librarians generate search keywords from key concepts is important; this knowledge is valuable both for researchers conducting the themselves, and for those developing automated tools for this purpose.

### 2.2.1 *Guidelines for the Use of AI Assistance*

Existing guidelines in general contexts recommend caution and transparency when using AI-assisted tools, for instance, Lo (Lo, 2023) introduces the Concise, Logical, Explicit, Adaptive Reflective (CLEAR) Framework. This framework helps educators guide students in creating effective prompts for AI language models like ChatGPT. By following CLEAR's principles, students can become more critical thinkers, better understand AI-generated content and leverage AI for research tasks. On the contrary, Acar (2024) proposed the PAIR framework to empower educators in proactively integrating generative AI tools into their curriculums. This framework centers on cultivating five key skills for responsible AI use: problem formulation, exploration of the AI tool, interaction through critical thinking and reflection. Students mainly follow a problem-solving framework where they define a challenge, select an AI tool, interact with it through prompts and evaluation and then reflect on their experience, similar to work proposed by Zheng *et al.* (Zheng *et al.*, 2024). However, there is a lack of specific guidance for researchers conducting literature reviews using computational tools (including AI) within the HCI field—much less for doing so in an ethical, comprehensive and replicable way. There is a need for guidance that addresses ethical considerations, potential biases within AI tools and the unique challenges of navigating the vast and nuanced body of academic literature, as advocated

[5] Systematic literature reviews use PICO to formulate research questions by identifying key elements for searching and evaluating studies effectively.

[6] NICE offers evidence-based guidelines and tools for conducting systematic reviews.

[7] SPIDER is an alternative to PICO suited for qualitative and mixed-methods research with a focus on study design and evaluation criteria.

[8] JBI is an international research organization providing comprehensive resources for healthcare reviews.

[9] PRESS is a guideline for enhancing the quality of literature search strategies through peer review to ensure thoroughness and accuracy.

by many authors (Bommasani *et al.*, 2021, Hancock *et al.*, 2020, Lund *et al.*, 2023, Safdar *et al.*, 2020, Stahl *et al.*, 2021)—we echo this for the context of HCI and HCI SLRs, specifically.

Traditional manual search methods are time-consuming, making the promise of automation through AI assistance tempting. This leads to an important question: what are the practical considerations when incorporating AI tools without established HCI-specific guidelines? While AI offers powerful assistance, a gap exists between its capabilities (which also still need to be more clearly established) and the specific needs of HCI research. This gap is further amplified by potential biases within some AI tools, often geared towards medical research. HCI researchers can benefit from the expertise of librarians who are well acquainted with HCI research methods. These librarians possess a deep understanding of best practices for conducting formalized literature reviews. Their information literacy skills and vast knowledge of information resources empower them to ensure a comprehensive search that encompasses all relevant sources, not just those readily accessible by AI tools (Corrall, 2012, Meert *et al.*, 2016, Spencer & Eldredge, 2018, Wagner *et al.*, 2022).

# 3 Methods

To understand current best practices for the search and selection stages in formalized literature reviews (RQ1a) and explore experts' perspectives on the role of computational tools in the search and selection stages (RQ1b), we recruited a purposive sample of eight librarians with expertise in systematic reviews and primarily those who have knowledge in HCI literature. Their knowledge of information science, search strategies and information literacy allows them to guide researchers in crafting comprehensive searches and critically evaluating sources, leading to rigorous and reliable results (Corrall, 2012). In addition, most librarians must stay updated on the latest information retrieval trends, which include computational tools for systematic reviews. They can offer valuable insights into the potential and limitations of these tools for the search and selection stages (Meert *et al.*, 2016). Thus, librarians' expertise positions them to provide guidance on both best practices for conducting rigorous reviews (RQ1a) and the role of computational tools within that process (RQ1b).

In-depth semi-structured interviews were conducted by the first author, with open-ended questions asked to explore the participants' thoughts, common workflows, understanding and concerns relating to formalized literature reviews, as well as the use of computational tools in this context. We chose interviews as they provide a flexible structure with predefined questions but also the exploration of emergent themes, facilitating in-depth insights and a nuanced understanding of librarians' perspectives. Our interview questions target the librarians' recommendations for conducting transparent and reliable multi-database searches, suggested criteria for screening the initially found papers for selection in the final corpus, and their thoughts on computational tools in these review stages. This study was approved by the Research Ethics Board at the University of Waterloo to follow ethical research guidelines. At the end of the study, participants were given a thank-you letter.

## 3.1 Recruitment and Participants

We initially recruited librarians through e-mails and social media posts. Librarians with expertise in systematic reviews can be a difficult demographic to locate, especially with experience in the HCI domain specifically. We also used snowball sampling to expand our reach and access a broader pool of qualified participants (Naderifar *et al.*, 2017) by asking initial participants to recommend colleagues with relevant experience in systematic reviews. Following up on these recommendations helped us recruit a more representative sample of librarians with the desired expertise. In total, eight librarians were recruited who possess a Master's degree in Library and Information Science (MLIS) or equivalent qualification and have published at least one systematic review and/or meta-analysis in the field of Human–Computer Interaction (HCI) in the past five years OR have experience with conducting systematic reviews and meta-analyses in HCI, either as a researcher, reviewer or editor. All have substantive expertise in the formalized literature review process (primarily SLRs). The interviews were expected to last 30 minutes; all participants voluntarily continued the interview for longer, leading to an average duration of 48:51 minutes Table 1.

## 3.2 Analysis Approach

The interview data was analyzed using the thematic analysis (TA) approach (Braun & Clarke, 2021). This was conducted primarily by the first author in consultation and discussion with the other authors. All authors are well-versed in SLRs within HCI, having conducted multiple SLRs. All authors were involved in developing the themes in multiple iterations using affinity clustering. The audio recordings were transcribed in *Dovetail*[10] and were edited by the first author to check for potential errors.

[10] Dovetail is an online software that enables users to extract valuable insights from interviews, surveys and feedback. By applying tags and codes to transcribed data, users can efficiently analyze information and generate comprehensive summaries.

**Table 1.** This table presents demographic information about the participants and their areas of specialization according to the CHI Subcommittee. Additionally, we note that librarians also provide instruction on conducting literature reviews and possess expertise in fields beyond HCI at their respective institutions.

| Participant | Gender | Years of Experience[1] | Specialization in HCI[2] | Interview Time |
|---|---|---|---|---|
| LIB1 | Male | 10+ | Information Visualization | 84:44 minutes |
| LIB2 | Male | 10+ | Computational Interaction | 57:36 minutes |
| LIB3 | Female | 20+ | Engineering Interactive Systems | 42:16 minutes |
| LIB4 | Female | 20+ | Health | 40:06 minutes |
| LIB5 | Female | 20+ | Health | 39:13 minutes |
| LIB6 | Female | 10+ | Education, Health | 36:40 minutes |
| LIB7 | Female | 10+ | Engineering Interactive Systems | 41:22 minutes |
| LIB8 | Female | 20+ | Accessibility Design | 48:51 minutes |

[1] Years of experience are categorized into buckets: '10+' represents 10 to 19 years, while '20+' indicates over 20 years. This categorization is used to maintain anonymity. [2] Specialization is based on the categories identified by the CHI Subcommittee, as self-reported by the librarians.

**Table 2.** A summary of our findings related to the research questions (RQs). We create the suffix at the end of each theme: 'a' in relation to RQ1**a** and 'b' in relation to RQ1**b**.

| Findings | Description |
|---|---|
| ● Theme 1 (T1a) | ⚓ Finding Your Anchor To Fish For Corpus Papers |
| ● Theme 2 (T2a) | A Hooking The Right Keywords |
| ● Theme 3 (T3a) | Casting Your Keywords In The Right Database |
| ● Theme 4 (T4a) | What Goes Into The Net (Corpus) And Knowing When to Stop Searching |
| ● Theme 5 (T5b) | Involvement with Computational Tools for Pre-Literature Review Brainstorming |
| ● Theme 6 (T6b) | Transparency in AI Relevancy Ranking |
| ● Theme 7 (T7b) | Foundational Knowledge: Prerequisite for Rigorous Research |
| ⚠ Unexpected[1] | Issues Revolving Around HCI Aspects |

[1] This was an unexpected finding that was not the focus of our research question.

We applied open coding to the transcripts. Any new relevant codes were merged when possible while keeping new entries open to interpretation under both flat coding and hierarchical coding framework models. Any codes irrelevant to the research question were labelled as *miscellaneous*. To capture the essence of each theme, affinity clustering of codes was conducted on a Miro board.[11] To generate these themes, we ensured they fit thematically (i.e. no contradictions). The themes were refined through multiple iterations; the final themes are presented in the findings below.

Our theme development focused on data directly related to the research question. The first level of clusters (not themes) was derived from the interview questions: Experience As A Librarian, Understanding The RQ, Researcher Involvement, Databases and Corpus, Search String, Criterion, Recommendations and Assisted Tools. A second level of categorical clusters (not themes) was derived from the process of conducting a formalized literature review, broken down into various stages. Focusing on the search and selection stages, we then created themes over several iterations based on short, self-explanatory sentences relevant to the research question that reflected the content captured in the different clusters (see Appendix for the full set of interview questions and an example of line-by-line coding).

## 4 Findings

In the first section, we present themes about what our experts (librarians) recommend for the search and selection stages in formalized literature reviews (RQ1a). This includes how this process is affected when focusing specifically on HCI-relevant databases. In the next section, we share themes relating to perspectives on computational tools within the search and selection stages of formalized literature reviews (RQ1b), including the feasibility of AI-assisted tools and concerns. We support each theme with participant quotes, using an abbreviation system where 'LIBX' stands for 'Expert HCI Librarian X' (e.g. ' LIB1') to indicate the source of a quote. A summary of our findings can be found in Table 2.

### 4.1 Build a Strong Foundation for Formalized Literature Reviews

Our experts consistently highlight the critical role of corpus acquisition in formalized literature review quality; strategies for this directly impact the representativeness and overall quality of the studies included in a review:

> *'The search is kind of the first thing, and I've always felt that the quality of your systematic review, or your lit search, really hinges on the quality of the lit search... you need to know that you've got all the literature out there'* ( LIB3)

#### 4.1.1 ● *Theme 1 (T1a): Finding Your Anchor To Fish For Corpus Papers*

Our experts noted that the process generally begins by identifying foundational work (seminal articles, seed papers or white papers)[12] relevant to the research topic. Ideally, the researchers' own *background knowledge* can guide review teams in selecting a suitable seed paper; alternatively this can draw on the expertise of working scholars or librarians, if the researchers are new to the research topic. However, it is important to consider potential background knowledge bias during this selection. Our experts emphasize that the purpose here is to establish a search strategy, not define inclusion and exclusion criteria for the literature review.

> *'one of the first steps I would encourage someone to do [is] the seed paper [...] , usually people or hopefully [people] have a sense of what databases they would like to search'* ( LIB4)

Our experts highlighted the importance of not only extracting references cited in those seed papers, but also exploring the *cited by* data. These two steps are known as backward and forward citation chasing or chaining, respectively (Hirt *et al.*, 2021). This can be particularly valuable if the seed paper is older, as it is likely to have been referenced more extensively:

> *'also look at the cited by data who has been citing that article because that might be really useful. Particularly if your seed article has been around for a while.'* ( LIB4)

One particular librarian stated that in HCI research, seed papers are particularly important because the field often produces less repetitive studies than other disciplines. This characteristic makes seed papers even more valuable as anchor points for the search strategy. By treating the seed paper as a guaranteed relevant result (a known 'hit'), researchers can fine-tune their search terms. This iterative process helps them develop a search string that consistently retrieves valuable studies:

[11] Miro is a digital collaboration platform designed to facilitate remote and distributed team communication and project management through whiteboarding.

[12] Seminal sources (also known as seed papers) are those that are first to present an idea of great importance or influence within a particular discipline, generally prompting many citations. A white paper is an informational document (generally issued by a company or not-for-profit organization) to reports on a particular topic, for example to promote or highlight the features of a solution, product or service.

*'If those articles don't come back, then you've done your search wrong.'* ( LIB7)

#### 4.1.2 *Theme 2 (T2a): Hooking The Right Keywords*

Librarians recommended that the first step after identifying your seminal paper(s) is to analyze their content and abstracts. This involves extracting key terms and the indexing vocabulary used by the authors. These terms will be the foundation of the search strategy. The focus on abstracts is because they typically provide a concise overview of the research and often serve as the first point of contact during the initial screening stage. Our experts explain that abstracts are also usually publicly available and not restricted by paywalls, making them a valuable and accessible resource for keyword extraction:

*'So you take your one (seed paper), normally you start the one that you do your search on is your primary database. The one that you think you're gonna get the most and best results from, and then you move along from that one, and you're always searching title, abstract, subjects and keywords.'* ( LIB5)

Our experts noted that having experience in the field can provide valuable insights into relevant indexing terms. For instance, keywords chosen by HCI authors for their papers can be inconsistent across databases, which can pose challenges for a formalized systematic review.

To address this inconsistency and broaden the search, our experts suggest leveraging previously published papers related to the targeted topic specifically as rich sources of additional keywords and relevant citations. As LIB4 comments, '*maybe they have a couple of papers already, and now you can mine those papers for keywords, or look at citations cited by citations.*'

Our experts recommend a 'cited-by/cited-in' search to refine the search strategy further. This involves analyzing the papers you find for keywords and then exploring the references those papers cite, and who cites them. By following this chain of citations, you can learn more about the foundational works that could be missed. However, the focus is to identify new keywords to expand the search.

Further, our experts highlight the importance of considering subject headings,[13] which are consistent tags or labels that describe what the item (book, article, etc.) refers to when searching databases relevant to HCI research. These standardized terms function as a controlled vocabulary, facilitating the retrieval of relevant information through precise categorization and organization of content. Unlike searches with keywords, subject headings eliminate ambiguity and allow for precise targeting. They are often organized hierarchically, enabling researchers to narrow down their search to specific subcategories within their research topic. This improved specificity translates to more relevant and efficient search results. For example, if this were more extensively implemented in HCI databases, a researcher could search under 'Game User Interfaces—Usability Testing' and reliably find all papers conducting or addressing usability testing in game user interfaces, instead of searching by those keywords and potentially missing results that talk about a specific game user interface without using the term 'game user interface':

*'Assigned subject headings are something you should keep track of. You, of course, can look at the author-supplied keywords that are often provided.'* ( LIB4)

*'Subject headings can be quite powerful if you tap into the right ones'* ( LIB8)

As one expert aptly noted, although online searching has replaced physical card catalogs, the core principles of subject headings remain unchanged. Some researchers still resort to manual searching, as not all digital resources are accessible online. Modern library and database services platforms pull metadata from various sources, including traditional high-quality library cataloging, publisher data and open-linked data (linked within the term subject heading). While the dominance of subject headings might have lessened with the influx of diverse metadata sources, they still play a significant role in influencing search results:

*'Our metadata, that we create is based on basically a card catalog, that's how it develops. So before online searching, there were these big cabinets and cards were typed out. I know it sounds like ancient history, but that is how our encoding scheme developed because that just got moved on to a record author title, publisher, and extent of the item subject headings just got moved into a mark record. So it does affect the way that you search, I mean a keyword search can pick up in any field, but a more precise search would be author, title, subject heading'* ( LIB8)

Lastly, our experts recommend conducting iterative searches with various keywords, including names and synonyms, to ensure comprehensive results:

*'You can't just do one search and forget about it. You've got to try like multiple times with multiple different terms, using keywords, using names, using different words. You're just not gonna find everything in the first search.'* ( LIB8)

#### 4.1.3 *Theme 3 (T3a): Casting Your Keywords In The Right Database*

Our experts emphasize that formalized reviews generally rely on peer-reviewed articles as gathered in various databases, although some researchers might resort to a quick Google Scholar search initially:

*'So in Google Scholar, there are no magical things. So it would be a very general search... Where some of the other databases you might have to tweak them because you don't wanna have 50,000 hits and have to go through 50,000 things when a lot of them are not as good as potentially other things.'* ( LIB1)

Experts recommend choosing a database that aligns with the appropriate field's research practices. In the case of HCI, they mention ACM Digital Library (ACM DL)[14] , Scopus[15] , IEEE Xplore[16] and more. However, other databases may be more important when specific topics are covered by different disciplines.

[13] In the ACM Digital Library, this would consist of the ACM Computing Classification System (CCS): https://dl.acm.org/ccs.

[14] The Association for Computing Machinery (ACM)'s Digital Library includes conference proceedings, journals, magazines and technical reports. It offers a strong selection of HCI research papers.

[15] A large citation database published by Elsevier that indexes a wide range of academic journals, conference proceedings and books.

[16] The Institute of Electrical and Electronics Engineers (IEEE) is a large organization for professionals in engineering, computer science and related fields. Its library offers access to a vast database of research papers, standards and other technical information.

In addition, databases can have inconsistencies with the search queries, as LIB1 notes: *'You can kind of fire out your search today and fire your search tomorrow and get potentially different numbers of hits and potentially different orders of hits.'*

Using an example provided by LIB6, when an HCI student approaches a librarian with a biology-related topic, the librarian would likely prioritize identifying relevant 'primary databases' in the field of biology-focused HCI (BioHCI). While multidisciplinary resources like Google Scholar can provide a valuable foundation, there are also specialized biology databases that would offer significantly more focused and relevant results for the student's research:

> *'If like a biology student were asking me a question and that's not my area of comfort. Usually, my mind would kind of go right to OK, this is a biology question. What are some of the key databases in biology? Like I've got these multidisciplinary resources that I know for sure are probably gonna be good; something like Google Scholar, Academic Search Premier, but then there's also biology specific databases that the student would need to know about. And that's the case for really any subject'* ( LIB6)

Experts suggested using Boolean operators strategically: combining keywords with AND to ensure all terms appear in the results, broadening the search via OR and excluding irrelevant terms via NOT. However, in HCI reviews, multiple different databases may be relevant, and string operators might not be directly transferable across multiple databases due to variations in syntax. These seemingly minor changes can significantly affect the size of the corpus. Our experts suggest always testing the search string with the seminal paper(s) again to ensure it captures the right materials. While basic functions like quotations and asterisks tend to function similarly, some platforms have their own unique syntax:

> *'So you have to kind of make sure that as you're switching, especially between like the host of the database... you just kind of translate over the phrases as closely as you can... a little bit of testing usually too... because it's sometimes the databases might have different pools of literature they're pulling from.'* ( LIB3)

Our experts emphasize that different types of literature serve different purposes, for example books usually offer foundational knowledge and definitions, while journal articles and conference papers provide current and in-depth research. Our experts noted that depending on the nature of the study, you might also need statistics, patents, standards, reports, theses, etc. It is important to understand that different databases specialize in different types of literature within HCI. For instance, some databases provide a list of indexed journals, allowing researchers to ensure coverage of specific papers. In addition, our experts recommend checking for adjacency operators when translating across databases, because these operators allow researchers to specify the proximity of terms within a search query:

> *'So you always have to change your subject terms because those are never the same across and then you have to look at what their adjacency operators are and that kind of thing and just make an equivalency.'* ( LIB5)

It's important to acknowledge that not all databases have the same indexing capabilities as medical databases. Every large database search engine has its own quirks in interpreting queries.

Further, even when two different individuals conduct research on the same topic, the approaches and resources they utilize can vary significantly. This causes search strategies to diverge:

> *'So if you're doing your research versus me doing your research, you might know about some grey literature bits and search there where I might not. So that's where our reviews themselves start to diverge.'* ( LIB1)

#### 4.1.4 ● *Theme 4 (T4a): What Goes Into The Net (Corpus) And Knowing When to Stop Searching*

Our experts caution against potential bias in HCI research that heavily favours Western literature. While systematic reviews often prioritize English-language sources for practicality, this approach can unintentionally exclude valuable research from other regions. Our experts note that HCI flourishes on diverse perspectives, and neglecting non-Western literature could lead to overlooking innovative approaches and insights. To combat this bias, our experts recommend incorporating multilingual search techniques, collaborating with international researchers and utilizing citation mining tools. These strategies can broaden the resource pool, offer access to local publications and reveal research across languages, ultimately leading to more inclusive reviews and a richer understanding of HCI on a global scale. While machine translation may not be perfect, existing translation tools can already be leveraged to some extent to aid in incorporating research from other languages:

> *'Another fun thing that comes into play is languages. So if we're doing a search in Scopus, all the materials, all of the abstracts are probably going to be in English. But there's a chance that the publication itself is not in English. So what do you do? So there are some transcription services available.'* ( LIB1)

> *'We tend to favour Western literature. So having some of that international access is good.... So they also have controls or index words that you can kind of like compare, like I would just kind of like be searching for the same words that I used in the initial search to translate over.'* ( LIB3)

Our experts also noted that the increase in size to the tentative corpus with each search iteration will vary. Once you reach a point of diminishing returns with your refined search terms, you're likely capturing the most relevant studies. Our experts often refer to this as a 'point of saturation'.

> *'the point of saturation is when you're not finding anything new; you just keep finding the same old things, no matter what you do, things you've already seen'* ( LIB7)

Systematic reviews—particularly in the medical field—are intended to be comprehensive, exhaustive and rigorous, but they can be time-consuming to complete. As LIB1 mentions, *'you're looking at probably a year plus to actually finish it.'* According to our experts, reviews conducted in less than one year typically fall under the category of rapid literature reviews, and may not be considered formal SLRs due to the time constraints involved in comprehensive searching and analysis. Additionally, if not published promptly, the findings may become outdated. To address this, our experts recommend updating your search before attempting to publish. By revisiting the literature periodically, you can ensure that your review remains current and relevant:

> *'then obviously you have to write it all up and you have to update your search before you try and publish because if this is the whole process is taking you a year, you have to update it to see if there's anything new that you need to add, which usually isn't as bad as, you know, the first ground searching and stuff, but it's still kind of an expectation that you have to do as well.'* ( LIB1)

As one final step before publication, they emphasize transparent reporting to ensure the same corpus resulting from replication:

> *'And as long as you describe each of your processes in sufficient detail, someone else should be able to do the same thing with the same information and get similar results, which is how we go from literature, which is light and fluffy to our scientific process..... If you go through the same stuff and follow the same procedures, you should have the same results.'* ( LIB1)

## 4.2 Impact of Automating Formalized Literature Reviews

In this section, we present the perspectives of our librarian experts on computational tools used for literature reviews.

### 4.2.1 ● *Theme 5 (T5b): Involvement with Computational Tools for Pre-Literature Review Brainstorming*

Our experts think that computational tools can help brainstorm ideas for exploring interesting topics before the literature review process, to help the researcher learn more about their research area prior to the actual review process, e.g.:

> *'I think that maybe there is like that potential for you to get new ideas of what other questions you could be asking, that may or may not match what you initially had thought would be the most interesting thing about the topic that you're exploring.'* ( LIB2)

For example, by generating potential research directions and uncovering emerging themes, these tools guide researchers toward relevant topics that may not have been initially considered. This capability streamlines the search process by narrowing down key areas of interest, making it easier for researchers to focus on specific topics. Our expert notes that insights into the most pertinent literature, these tools can help researchers to explore and evaluate the most relevant sources with respect to their research goals.

However, in the context of exploration, researchers may use AI tools that require them to upload papers downloaded from their university library. Our experts advise caution when using documents obtained through the university's VPN access in this way. These Portable Document Format (PDF) often come with digital rights management (DRM) or contain a university watermark that can violate the interlibrary loan (ILL) policies.[17] Our LIB8 notes, *'many are unaware; while laws are not in place yet, they have serious issues down the road'*. Our expert notes these measures are designed to prevent unauthorized distribution and sharing.

### 4.2.2 ● *Theme 6 (T6b): Transparency in AI Relevancy Ranking*

Our experts raised strong concerns about the 'AI blackbox' phenomenon and how it applies to search algorithms. This lack of transparency makes it difficult to assess the relevance and reliability of retrieved information. Since the factors influencing search results remain unclear, automating tasks like literature review searches with these tools can be problematic:

> *'But there's always been this black box, at least for me, I don't think that any company is 100 percent transparent when it comes to the way that your results are outputted. When you start with relevant, what relevance means is a mystery. So when you sort by title, it's alphabetical. When you sort by author, it's alphabetical. When you sort by subject, it's alphabetical, it's not unbiased, but there is a method to the madness when it comes to relevance ranking and even in traditional database searches.'* ( LIB2)

### 4.2.3 ● *Theme 7 (T7b): Foundational Knowledge: Prerequisite for Rigorous Research*

> *'You get that from understanding your field and from working in your field for many, many, many years, because you wouldn't be able to know that as an undergraduate student.'* ( LIB2)

Our experts have observed a delicate balance between the desire for rapid solutions, especially among undergraduate researchers, and the necessity for a nuanced comprehension of the research question. They emphasize that relying solely on AI tools without critical evaluation can result in superficial searches, potentially missing relevant or unbiased sources. Even when the inner workings of an AI system are transparent, researchers with substantial background knowledge can contextualize search results more effectively and critically assess any biases introduced by the search tool or their own prompt formulation. In cases where transparency is lacking, experts recommend human validation to verify results:

> *'you still need to go in there and make that final decision.'* ( LIB1)

As one librarian expressed, while undergraduate students may not possess the experience needed to contextualize AI-generated results fully, seasoned researchers (including senior PhDs and professors) often develop an intuitive understanding of their field. Even when faced with results from AI black boxes, this expertise allows them to gauge whether the outcomes align well with the current state of knowledge:

> *'So not when you're an undergraduate, but maybe you're a PhD, you're a professor, even if you don't know what the black box is throwing at you, you have a sense for this is where my field is at. And so, I think that we will need a lot more of that. And you don't get that from using the AI tools. You get that from understanding your field and from working in your field for many, many, many years, because you wouldn't be able to know that as an undergraduate student.'* ( LIB2)

> *'There's a process that we call the "reference interview" which begins kind of when somebody asks you a reference question. Rather than just kind of taking that question at face value, you wanna ask more follow-up questions to get more details and to figure out, like, what exactly is this person looking for?'* ( LIB6)

Our experts emphasize the importance of a process known as the *reference interview*. This interview goes beyond simply responding to a researcher's initial question at face value. Instead, librarians engage in a focused dialogue by asking clarifying questions. This deeper exploration helps to uncover the researcher's

[17] ILL is a cooperative arrangement between libraries by which they agree to share their collections with each other.

specific needs and research goals. By understanding the researcher's intent, the librarian can tailor their support and provide the most relevant information resources. However, LIB1 articulates: '*Undergrads and librarians don't talk; however, when you go for your master's or your PhD, you might be able to talk to a librarian.*'

Lastly, our experts highlight the possibility that existing biases in the information landscape could be significantly amplified when AI automation in reviews is used increasingly. In certain conversations, relevant voices may remain hidden because they did not use the specific keywords that algorithms seek or originate from sources outside databases used for training. Thus, these technologies can create an illusion of objectivity:

> *'The algorithm is looking for, or they don't come from, the publications that are within the databases that are being trained. And so, sometimes, these technologies can give the appearance of objectivity.'* (LIB2)

### 4.3 Issues Revolving Around HCI Aspects

Our research question focused on understanding the factors influencing student satisfaction with online courses. While analyzing our data, we unexpectedly discovered a recurring pattern related to database inconsistencies and issues within HCI-formalized literature reviews.

To further iterate transparency, we report the questions that revolved around this unexpected pattern; '*How do you determine that the searches are comparable across all databases you are using in the specific review?*' and '*What do you recommend for researchers about how to maintain consistency and reliability in the screening process, especially with multiple reviewers?*'

> *'We did a search one day, and the exact same search, even a couple of hours later, was returning different numbers of hits, which in my mind says it's a crap database (ACM). I don't know if they weren't able to access the same items the second time or what was going on. But I know the search was the same because I copied and pasted it from the same document and put it into the same search areas, and the results were not consistent, which would make it very difficult for someone else to replicate it.'* (LIB1)

Our experts point out that the ACM database showed significant inconsistencies, giving different search results for the same queries within a short time. This makes research less reliable and harder to replicate. Also, the absence of a feature to save searches complicates managing research and tracking progress. These problems suggest the database should be improved for consistent search outputs.

> *'The problem, (heavy sigh), journal articles, right? That you have a limit, you can't go beyond it these days. Sometimes you're able to have more like supplements and stuff on, on the website. But yeah, it's hard to like condense the search, and so often, like, that's not the part people are interested in either. So it's not, it's often not pushed to be like, published fully.'* (LIB6)

Our experts note that journal articles often have strict word limits, forcing researchers to condense their search strategies for publication. Although supplementary materials and external websites provide options for sharing more detailed search information, these additional resources may not be widely accessed by readers or have the same level of importance as the main text. As a result, the important process of developing search strategies is frequently underrepresented in published research.

## 5 Discussion

In this section, we discuss the implications surrounding our experts' responses. We use librarians' expertise to find challenges they see with HCI databases, potential mismatches between their recommended best practices and what they and we (anecdotally) observe in HCI reviews, and the role of computational tools.

### 5.1 Librarian Recommendations vs. Common HCI Practice

As also noted by our experts, and observed in our own experience, there appears to be a gap between the best practices recommended by librarians and other guidance, and what is actually reported in HCI review papers. For instance, few SLRs in HCI mention a step in which they identify foundational work as seed or anchor papers (Rogers *et al.*, 2024). However, it is certainly possible that this step is being *done* and simply omitted due to length constraints.

Similarly, the truly expansive search expected from SLRs in other fields—including preprints, grey literature and sources in other languages—is quite rare in HCI (Benzies *et al.*, 2006, Paez, 2017). Limited resources, of course, play a substantive role in these issues, e.g. (Neimann Rasmussen & Montgomery, 2018) suggest that this factor might hinder the inclusion of research in other languages. Additionally, the push to publish quickly and often in HCI is strong. However, we also note that given the diversity of HCI research, there is also simply a lot of secondary research being done that *does not have the goal of being exhaustive*, but rather aims for purposeful, representative sampling, and this can also be a valid approach to SLRs.

Interestingly, in HCI we may want to additionally consider adding artefacts or novel systems and any documentation of design decisions, design rationale and how the artefact is being used (e.g. in the form of websites or videos). We have not seen this kind of record being added to an HCI SLR yet, and it is unclear how this kind of knowledge should be synthesized, but this is likely something our field is well positioned to explore.

Finally, we note a mismatch in the recommendation for we note a mismatch in the recommendation of using subject headings; this currently has limitations in HCI. For example, in the ACM Digital Library, the current system consists of the 2012 ACM Computing Classification System (CCS), which is both rather old and also consists of broad categories. For example, the classifier '*CCS → Human-centered computing → Interaction design → Interaction design process and methods → User interface design*' might group together studies on very specific user interface elements like website buttons, but also papers discussing the design rationale behind tangible design artefacts, as well as papers broadly discussing design methods for interactive systems.

### 5.2 Varying Expectations of Time(liness)

There appears to be a discrepancy between commonly reported and observed durations of SLRs in HCI and what our experts consider the standard timeframe for such a rigorous process (minimum one year). It must be considered that this could indicate that parts of the expected steps or practices are not been completed or not being completed fully. It could also mean that many SLRs in HCI may actually be closer in form to a rapid review. Rapid reviews are marked by an accelerated process based on completing SLR steps with variations that take less time (e.g. omitting critical appraisal or conducting single rather than dual screening) (Hamel *et al.*, 2021). In our experience, such characteristics are quite common in HCI.

In line with Hamel *et al.* (Hamel *et al.*, 2021) suggestion, we recommend clear differentiation between systematic and rapid reviews—this first necessitates a community-led agreement on developing or choosing clearer definitions or criteria for these different types of reviews. In addition to expectations about duration, the interviews with the librarians suggest that HCI as a field may harbour different (and likely a variety of) expectations regarding the timeliness of reviews, specifically when or how quickly reviews become outdated. In fields where meta-analyses (which calculate an overall effect size for interventional research questions across all available studies), missing a very recent study can have a big impact on the final outcome, as it can change the primary desired outcome (the overall effect size). In HCI, the pace of research publication is similarly fast. But if the review does not focus on an interventional research question but rather is of a broader nature, then missing a single (or a few) papers is unlikely to change the answer to the research question substantially.[18]

An alternative solution to the potential issue of timeliness could be living systematic reviews, which aim to provide continuously updated summaries of the evidence (Elliott *et al.*, 2014). HCI, in particular, may be well positioned to develop such living reviews as interactive systems (Rogers & Seaborn, 2023), though it, of course, entails maintenance challenges.

## 5.3 Who Should Be Conducting HCI Reviews?

Current research suggests that systematic review teams benefit from including members with extensive expertise in the reviewed topic area (Cooper *et al.*, 2018), which may be less pronounced in very junior researchers. Weber *et al.* (2019) also point out that strong information literacy skills are lacking in undergraduates. Librarians may thus be a valuable resource for helping more junior researchers develop their search strategies. Unfortunately, our experts point out that undergraduate students rarely seek librarian assistance, and even graduate students are hesitant to do so. Anecdotally, we have observed that young PhD students in HCI often begin review papers as a first (or early) research project. Yet we have only rarely heard of them looking for help from librarians or information specialists, or being given methodological guidance from more experienced researchers with both domain knowledge and formal training in review methods (for example, to advise the student on appropriate methods, e.g. the choice between a systematic review and a rapid review). For the students, it can be a very valuable learning experience, but we do think it is worth discussing as a field whether our knowledge base would change if reviews were increasingly conducted with heavy involvement from more experienced researchers.

### 5.3.1 *Saturation: Depth vs. Breadth*

Librarians particularly excel at achieving comprehensive breadth because they are accustomed to navigating a wide range of topics and sources. Their expertise in search strategies can complement researchers' domain knowledge, hopefully leading to more focused and efficient searches that capture the most relevant studies, and can also help researchers identify studies outside their usual search patterns that might otherwise be missed.

We note that the librarians in our interviews conduct collaborative 'reference interviews' with researchers who reach out for advice. These are meant for librarians to gain insight into researchers' specific needs and research goals. While (our) librarians may be more used to breadth-focused approaches, such interviews could be used to indicate a review's focus on depth, instead, in which case the librarians can tailor their support for that purpose.

## 5.4 Benefits and Dangers of Automation & Assistance

Our interviews reflect a great deal of caution on the part of the librarians when it comes to computational tools in SLRs. For most of them, AI tools are only useful and reliable when it comes to brainstorming relevant keywords, rather than for actual search. The HCI and AI fields would likely generally agree that transparency, explainability and replicability are important factors in AI applications (and their impact and adoption) (Gundersen & Kjensmo, 2018, Samek & Müller, 2019, Tsafnat *et al.*, 2013, 2014). A similar note of caution is also reflected in interview results reported by Wu *et al.* (Wu *et al.*, 2024) based on PhD students from several fields regarding other aspects of academic search, e.g. AI-provided scores on papers' study replicability. However, this has not been extensively explored for the search results themselves—this is particularly important in HCI, given the issues described in our main databases.

Finally, as mentioned in our findings (**T1b**), some institutions have watermarks embedded on articles when they are accessed through a university's virtual private network (VPN), which can be problematic when using computational tools for search exploration (e.g. uploading the paper to extract suggested similar keywords). These watermarks serve as a form of digital rights management, indicating that the article was accessed through a specific institution. Watermarked articles in the cloud could raise privacy and data security concerns and could be a complex future issue for policymakers as they touch on issues of intellectual property rights, data privacy and academic integrity. Policymakers would need to navigate these concerns while ensuring the advancement of academic research and technological innovation. This includes the need for institutions to consider open research practices to facilitate literature searches and address the challenges identified above, such as watermarked articles. We argue that computational tools should prioritize the development of transparently explaining their decision-making processes to address the 'AI Blackbox' mentioned in **T2b**.

In addition, to mitigate biases arising from unequal access to literature, these tools should incorporate strategies to address the challenges posed by proprietary databases and watermarked content. Promoting open research practices and ensuring fair use of training data is important for developing equitable and effective AI-driven computational tools.

## 5.5 Concerns with HCI-Related Databases

Formalized literature reviews face a significant challenge when there are inconsistencies in how databases handle search queries (for example, whether they allow wild card search functionality), especially when the goals include replicability of the search. For example, the ACM Digital Library does not support adjacency operators (such as NEAR, WITH, etc.) that would allow researchers to search for keywords in specified proximity to each other. Changes to databases' functionality could improve the search power (allowing more fine-tuned searches) and better generalizability across databases. Bailey *et al.* (2007) have similarly highlighted a lack of overlap and many inconsistencies between key database search engines (including ACM Digital Library and IEEE Xplore). They also suggest researchers cannot rely on a single

[18] Consider a question of 'What are the most common factors causing accessibility challenges in interactive systems?' An answer describing the most commonly mentioned factors is unlikely to change dramatically when the corpus changes by one paper.

**Table 3.** The initial INSPIRE Framework describes and discusses key segments of the search process for literature reviews in HCI.

| Segment | Relevant Theme | Description |
|---|---|---|
| **[I]nquiry** | Theme 2 (T2a), Theme 6 (T6b) Theme 7 (T7b) | Inquiry about the intended research, assembling a team with complementary expertise and formulating clear research goals. |
| **[N]avigation** | Theme 1 (T1a) Theme 3 (T3a) Theme 6 (T6b) | Navigating the most appropriate databases and other resources to search for relevant literature, based on the identification of seed or anchor papers. |
| **[S]earch** | Theme 2 (T2a) Theme 3 (T3a) | Using a combination of convergence searching and divergent searching to identify relevant studies. |
| **[P]oint of Saturation** | Theme 4 (T4a) Theme 6 (T6b) | Find the point of saturation in the large volume of the literature identified in the search stage for relevance and quality. |
| **[I]nvolvement** | Theme 1 (T1a), Theme 5 (T5b) | Involving multiple researchers in the process to minimize bias and enhance the reliability of the review findings, and cautiously consider computational (including AI) tools. |
| **[R]eporting** | Theme 2 (T2a) Theme 4 (T4a), Theme 6 (T6b) | Reporting the search stage to ensure the methodology can be replicated. |
| **[E]xhaustion** | Theme 2 (T2a), Theme 4 (T4a) | Making sure the search remains as exhaustive as possible by scheduling periodic revisits to the search terms. |

search database, but rather must use multiple for coverage. (We note that this may depend on the reviews' goal, i.e. breadth vs. depth, which we discuss further below.)

Liang *et al.* (2020) identified personalized search as a key desired feature for the ACM Digital Library in a survey of 157 users. The respondents (42–54%) wanted rankings of search results to be personalized based on their browsing, search history, click history or research interests. Yet this kind of personalization could hinder comparability in SLRs, and so we propose that databases like the ACM Digital Library should offer both standardized search options (for reliable, repeatable results) as well as personalized options for informal browsing or for exploring the field in non-formalized contexts.

Another potential feature that could address inconsistencies and support the rigorous documentation the librarians advocate for would be a time-stamped corpus system. This would allow researchers to download snapshots of search results for a database at specific times, allowing researchers to easily retrieve and compare past searches even when database content changes. In summary, quoting Wu *et al.* (2024): 'there is a growing need for more sophisticated, flexible, and intelligent tools' in the context of academic search.

## 6 INSPIRE Framework

Based on the interviews with the expert librarians, we developed an initial framework designed to optimize search and selection in formalized literature reviews by strategically integrating human expertise and digital tools. It contains seven segments: **I**nquiry, **N**avigation, **S**earch, **P**oint of Saturation, **I**nvolvement, **R**eporting, **E**xhaustion (and is thus named INSPIRE). The segments are not intended to follow a clean-cut sequential order; rather, they should be understood as an iterative and often parallel process in which the segments may overlap, as well as be paused and be returned to. Table 3 describes how each of the seven segments in the INSPIRE framework is derived from the themes developed in our interview analysis while Appendix–Fig. A4 visually represents the framework's application and evolution over time.

This framework is ideal for use during the planning and execution stages, where researchers need to develop and document transparent, rigorous and reproducible search and selection strategies. However, it may not be suitable for reviews requiring strict adherence to standardized protocols from other disciplines, such as medical research. The framework supports repeated cycles of searching, validation and refinement and therefore enables researchers to improve the depth and relevance of their search process iteratively.

**Why Is There A Need For This Framework?** When conducting a formalized literature review, the best practice is to clearly document the process (**T1a**). However, within HCI, the specific processes for conducting such searches remain unclear for literature reviews (Rogers & Seaborn, 2023, Stefanidi *et al.*, 2023). Most often, HCI researchers implicitly rely on methodologies derived from medical literature searches (e.g. PRISMA), which may not fully encompass the diverse range of research strategies employed within HCI papers, and cannot draw on attempts of standardized terminology such as through subject headings.

While work by Gusenbauer (2022) has explored disciplinary coverage—the extent to which literature of different fields is prevalent—in common databases, the coverage of HCI topics remains an open question. Further, the ACM Digital Library (as our field's likely most relevant database) has noted search reliability issues and is poorly documented (Rogers & Seaborn, 2023). This demonstrates a knowledge gap in how existing databases and computational tools can or should be used in the search and selection stages of formalized literature reviews. However, we do not see databases making drastic changes immediately. The lack of formalized guidance can hinder transparency and rigour, thereby limiting the reproducibility of search and selection stages. Often, HCI researchers implicitly rely on methods and strategies designed for medical literature searches (Atkinson & Cipriani, 2018), which may not fully capture the interdisciplinary nature of HCI research (Rogers & Seaborn, 2023, Stefanidi *et al.*, 2023). Furthermore, while AI holds promise as an assistive tool in the literature review process (Khraisha *et al.*, 2024, Mahmoudi *et al.*, 2024, Van Dijk *et al.*, 2023), there is a lack of clear guidelines on how to adapt search strategies to leverage AI for both search and selection stages effectively but also cautiously. Thus, establishing standardized approaches specific to HCI research could significantly enhance transparency and facilitate the replication of research findings.

## 6.1 Inquiry: Learn About The Research Space

The first segment involves inquiring about the intended research. The literature review team should ideally determine whether they possess relevant background knowledge or expertise in the field. This aligns with best practices in systematic reviews, as highlighted by (Kugley *et al.*, 2016, Lefebvre *et al.*, 2008, Petticrew & Roberts, 2008), because HCI encompasses various research domains.

For a well-conducted systematic literature review, assembling a team with complementary expertise is optimal, as noted in our findings. Researcher familiarity with the specific HCI topic allows for the development of more targeted search strings and terminology, improving the efficiency of the search and selection phase, as noted in **T2b**. Ideally, the team should comprise researchers with relevant subject matter knowledge *and* an information specialist (such as a librarian) (Reviews, 2009, Meert *et al.*, 2016, Okoli, 2015, Spencer & Eldredge, 2018). Information specialists bring valuable experience in knowledge synthesis and information retrieval, which are critical for executing comprehensive and efficient searches to support the inquiry process (Desmeules *et al.*, 2016, Spencer & Eldredge, 2018). Additionally, post-secondary institutions typically participate in ILL networks, providing access to a wider range of databases. Limited access to ILL networks in specific universities can hinder review conduct and replication (Boucher, 1997). For some databases, researchers may have access to abstracts and titles, but not to the full articles, leading to article exclusions in systematic reviews. An information specialist can assist researchers in navigating ILL procedures to identify alternative access options, minimizing article exclusions.

Further, this segment involves formulating clear and concise research goals that the review aims to answer (Meert *et al.*, 2016). These research goals serve as the foundation for the subsequent stages of the review process, influencing the search strategy, selection criteria and data extraction methods in line with other search and selection strategies (Cooper *et al.*, 2018, 2019). For instance, Cooper *et al.* (2018) highlighted preparation as one of the eight key stages in literature searching in SLRs in medical fields. This includes determining if there are any existing or ongoing reviews or if a new review is in progress, which is not always transparent in HCI. One can view this as a pre-requisite segment prior to conducting a formalized literature review.

## 6.2 Navigation: Find the Seed / Anchor Papers

The second segment titled navigation, should commence with the identification of foundational work (Herrmannova *et al.*, 2018). These works are relevant to the research topic and can act as the starting point (anchor or seed) for the literature review.

In this phase, researchers should take on the role of *navigators*, who map out the space of relevant research. The success of this segment heavily depends on the identification of the most fitting databases for further pertinent literature (Jalali & Wohlin, 2012, Smucker & Allan, 2006). Researchers must leverage their expertise to target the databases and resources most likely to yield relevant catches. For instance, if an identified seed article originates from IEEE, then including IEEE as a primary database would be prudent. Furthermore, considering where the article is indexed is important for comprehensive searching, as it might be cross-referenced on other platforms. In line with our findings, work by Brocke *et al.* (2009) suggests the selection of databases should include a thorough review of each considered database's search functionalities and coverage.

Both research above (Brocke *et al.*, 2009, Jalali & Wohlin, 2012) highlight the importance of these activities in the navigation segment of the literature review search process, the need for understanding the research domain and the significance of making informed decisions in this process. Our experts similarly highlight the iterative nature of the process, emphasizing the need for multiple rounds of refinement and exploration (**T1c**). While finding a good starting point (seed paper) helps to build a search strategy, it is important not to confuse this with setting strict rules for what studies to include or exclude in the review. The seed paper is a way to get the search rolling, not a way to decide what makes the final cut. Lastly, the searches for seed papers may turn up new potential seed papers, including potential ones published before the initial set. In this case, the new paper can be added to the seed paper set to *refine* the seed paper selection.

## 6.3 Searching: Different Approaches Of Finding Your Papers

The third segment comprises search, in which researchers should employ both manual and automated search strategies to locate potential studies for inclusion in the review. This necessitates the development of the actual search strings. This process in turn begins with the analysis of seed papers' content and particularly their abstracts, extracting key terms and indexing vocabulary used by the authors. As mentioned, our experts noted that these terms form the foundation of the search strategy.

This process always consists of iterative searches and checking whether the seed papers turn up. However, based on the experts' descriptions, this can be done with two different search motivations in mind, which we call *convergent searching* and *divergent searching*. We argue that a combination and careful but flexible iteration of keyword and keystring refinement via divergent and convergent searching is particularly relevant for literature reviews within HCI because HCI research is conducted by people from many different disciplines who use different terminology and publish in many different publication venues/databases.

### 6.3.1 *Divergent Searching*

When searching with this motivation in mind, the goal is to broaden the search and identify and explore new keywords, keystrings and databases for relevant research. Researchers use the initial research question and seed papers to identify additional related concepts, synonyms and alternative keywords that might lead to additional studies they might have missed. Other researchers like Ibrahim *et al.* (2024) even suggest to 'seek feedback from other researchers to help select unique, community-supported search terms'. If no or only a few seed papers exist, then divergent searching can help assist in finding some or more. It may even help to refine and iterate on the research question, as new approaches to a particular topic are found in unexpected papers. It ensures that the review does not overlook potentially relevant studies that might fall outside the immediate focus. Based on existing research, this process aims to broadly comprehend and explore a specific concept or body of scholarly work, considering its defining features (such as terminology, volume of evidence and research types) (Athukorala *et al.*, 2016, Gusenbauer & Haddaway, 2021, Marchionini, 2006, White & Roth, 2009).

**Example case:** Let us consider a researcher who is conducting an SLR on the topic of physiological measures in gaming. With divergent searching in mind, the researcher would start with the seed papers (perhaps a key study in the field that investigates how physiological measures like heart rate or skin conductance are used to evaluate player experience in video games). From this seed paper, they would extract potential new key terms, concepts

and methodologies used in the study. These might include specific physiological measures (e.g. heart rate variability, galvanic skin response), types of games studied (e.g. mobile games, virtual reality games) and any specific theories or models referenced (e.g. flow theory, player experience of need satisfaction model). They would then use these terms to refine the search, focusing on databases and journals where such studies are likely to be published. For instance, if the seed paper were published in the 'International Journal of Human–Computer Studies', it would be prudent to search this journal for more relevant studies. Using the newly identified keywords, they can test new searches to find more papers, allowing them to refine and add to the search terms and keystrings based on the new information gathered. For example, they might discover that some researchers used the term 'biometrics' instead of 'physiological measures', so they would add this to their list of search keywords and keystrings.

#### 6.3.2 *Convergent Searching*

This approach focuses on narrowing down the search strategy to maximize highly relevant papers and minimize irrelevant ones. With convergent searching as the motivation, researchers still leverage the research question, seed papers (previously identified above), initial keywords and keystrings to refine your search terms. However, the goal now is to determine which new keywords or aspects of the keystring are irrelevant (and can be omitted) or lead to not finding seed papers or finding overly large amounts of irrelevant papers.

Convergent searching aligns with Kraus *et al.* (2022) concept of 'focus' in literature reviews. By focusing on studies with high relevancy (resulting from a search with high convergence), researchers can gain a clearer picture of the established knowledge in the field.

**Example case:** Returning to the example of searching for papers on physiological measures in gaming. With the goal of convergent searching, researchers would still check for the inclusion of seed papers, but attempt to determine which keywords are actually unnecessary or add too much noise to the results. For example, the keyword heart rate variability might have seemed relevant at first, yet on further examination, the researchers might find that this keyword leads to many irrelevant results, while relevant ones seem to be covered by the already included umbrella keywords of physiological measures and biometrics. Alternatively, perhaps new keywords or different keystring operators need to be added to narrow down the result further.

#### 6.3.3 *Applying this to the Double Diamond Approach*

Iterative searching is particularly valuable in the field because it allows for continuous refinement and expansion of the search strategy across multiple rounds. Work by Allen (2000) highlights the complexities of applying individual differences research to user-centred design in information systems, particularly in the context of search-related tasks. Therefore, the literature review process can be conceptualized using the double diamond framework, a design thinking approach commonly used to structure problem-solving processes (Banathy, 1996). The double diamond reflects the iterative nature of the literature review, with convergent searching and divergent searching phases flexibly and iteratively occurring in each *diamond*, see Fig. 1. The first diamond focuses on 💎 '**designing the right keywords**' and the second diamond focuses on 💎 '**designing the keystring right**'.

We further borrow design thinking's discover, define, develop and deliver phases. In the Discover phase, the focus is on broadening the search (diverging) to identify a wide range of potentially relevant studies and, in particular, potentially relevant keywords. Researchers can achieve this by starting with the research question and (if available) seed papers to explore synonyms, related concepts and alternative keywords.

The Define phase centers on refining the search strategy based on the initial findings from the Discover phase. Here, researchers can analyze their initial results to identify and exclude irrelevant or overly broad keywords and concepts. This allows for narrowing down (converging) to a more focused set of relevant keywords. These two steps of the first diamond should be iterated flexibly multiple times until the researchers are satisfied with the search results.

In the Develop phase, the researchers can use the refined keywords identified in the Define phase to construct precise keystrings by combining the keywords with suitable operators. The keywords are combined with Boolean operators (AND, OR, NOT) and other search filters offered by the chosen databases (as mentioned above) to create a keystring. The results of the tested keystring that is the most promising at any time act as the baseline for the search query; it should be documented, and the results of new keystrings should be compared to it.

In the Deliver phase, researchers refine the keystring in the chosen academic databases, with the objective of retrieving the most relevant studies through comparable searches that form the foundation for the literature review.

We advocate that researchers should generously refine their approach. Tested keystrings should be documented to see how variations impact the retrieved literature (which eventually forms the corpus). Tracking the number of articles retrieved after each modification, and whether seed papers were found, helps visualize the effectiveness of the search strategy and identify if it is becoming too narrow or overly broad. Finfgeld-Connett & Johnson (2013) similarly highlight the importance of carefully iterating on concepts and their relationships in knowledge-building qualitative reviews, which aligns well with our own approach.

#### 6.3.4 *Citation Chasing or Chaining*

Our experts recommend a 'cited-by/cited-in' search to ensure a comprehensive review. We argue that this should be used to find an appropriate set of keywords using the double-diamond approach. This involves analyzing the papers you find for keywords and then exploring the references those papers cite and who cited them for the same purpose. By following the chain of citations, researchers can uncover foundational works they may otherwise miss. Tools like CoCites developed by Janssens *et al.* (2020) offer an efficient and reportedly reliable way to identify relevant and connected articles, though to our knowledge this has not been tested.

### 6.4 Point of Saturation

The fourth segment aims to determine a point of saturation; determining when enough papers have been found to finalize the search strategy and the corpus. The librarians clearly tend towards aiming for highly comprehensive searches. However, this may not suit all types of reviews (depth-focused reviews may instead use purposive sampling). Nevertheless, the iterative process of refining the search string requires weighing the papers found in terms of their overall fit and comprehensiveness in answering the research question.

Furthermore, database selection requires considerations of coverage, particularly for recently launched journals or those undergoing transitions. These factors can lead to missing valuable studies, especially in fast-moving fields like AI. Research

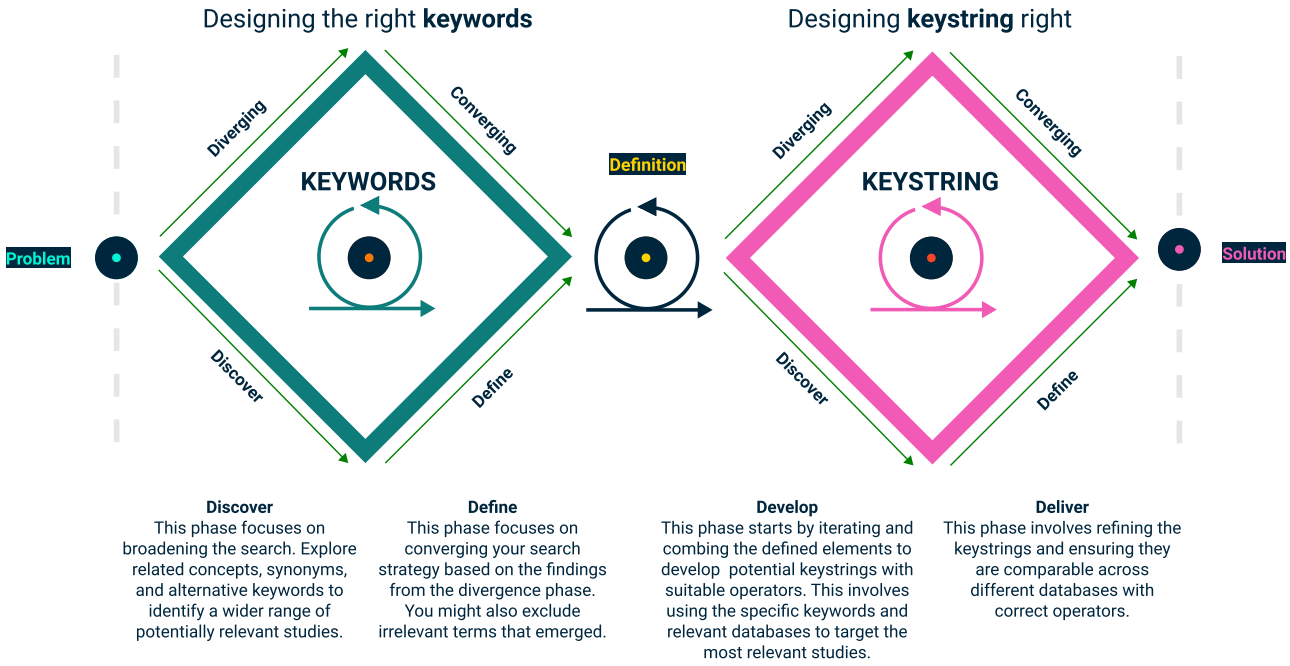


**Fig. 1.** This double diamond diagram illustrates the envisioned flexible and iterative search process in the INSPIRE framework.

by Wohlin *et al.* (2022) supports our recommendation that researchers conducting SLRs should give preference to databases with wide-ranging coverage. They also recommend incorporating multiple databases to lessen the potential for bias and guarantee a strong foundation of literature, which we echo. Our experts recommend refining the search terms iteratively until reaching a point of diminishing returns, also known as the 'point of saturation.' This indicates that the most relevant studies based on your initial focus have likely been found. Our experts noted that there is no *definitive number* of iterations to reach this point. Researchers sometimes reduce the scope of the years searched to make the corpus more manageable. While this can be a valid strategy, it should not be done solely for the sake of saving time.

## 6.5 Involvement: Inclusive Research Participation

Our interviews demonstrate that librarians specialized in HCI and information science can aid literature reviews by helping researchers identify pertinent databases and terminology (Rubin & Rubin, 2020). This resource should be considered by HCI researchers conducting SLRs. The involvement of librarians and information specialists becomes even more important when SLRs are conducted by novices, e.g. (under-)graduate students or junior students, as they can draw on their experience in information retrieval. This should be supplemented with experienced researchers from the specific review topic whenever possible. The segment suggests that novice reviewers should ideally work alongside experienced researchers and information specialists to benefit from their expertise in information retrieval.

### 6.5.1 *Using Computational Tools*

AI tools can search across numerous databases, repositories, journals and other sources simultaneously, ensuring a more comprehensive review of relevant literature than may be possible via manual searches alone (Khraisha *et al.*, 2024, Mahmoudi *et al.*, 2024, Van Dijk *et al.*, 2023). However, it is evident that the involvement of AI tools is not transparent (Beller, 2018, Khalil *et al.*, 2022, Tsafnat *et al.*, 2014, Schoot, 2021). At the current state of these tools, our experts suggest these tools should only be used for brainstorming until the AI Blackbox is transparent enough to allow replication. Many of these automated tools pull from public application programming interface (API) sources such as Semantic Scholar and Scholarly Py; these APIs allow tools to access scholarly databases for the purposes of automating searches or compiling custom datasets. This limits the documents that can be found.

## 6.6 Reporting: Document The Process

The reporting component is critical to the search process, as also suggested by others (Finfgeld-Connett & Johnson, 2013). It involves meticulously documenting the iterative search process and findings in a manner that is both clear and comprehensive.

Researchers should include their documentation of how they developed the search strategy in the supplementary material (to mitigate issues with page or word limits). This not only ensures the replicability of the literature review but also allows reviewers to follow the process in depth. The rationale behind the choice of key terminology should also be clearly explained. This is important because it provides context and justification for the chosen search strings, which are often 'locked' at the beginning of the project. This level of transparency is essential for maintaining the rigour of the literature review process and allows for constructive feedback and critique. It underscores the importance of justification in the search. It is highly recommended that researchers provide clear documentation of how the seed paper selection, keyword set and keystring variations evolved over time (Gross & Taylor, 2005). We also recommend summarizing how many papers were found for any main keystring milestones in the iterative process. Of course, the final corpus should be summarized in detail in the supplementary data. Such transparency not only enhances the replicability of the study but also allows for a more thorough review and critique of the research process. We recommend a flexible but careful cycling through the double-diamond approach with both divergent and convergent searching.

#### 6.6.1 *Reporting the Use of Computational Tools*

When researchers use computational AI tools in their SLR search, it is important that they document a clear understanding of how these tools function. This includes knowledge of the algorithms used, the data sources accessed and the criteria for ranking and sorting results. Without this understanding, other researchers may not be able to evaluate the effectiveness and reliability of the tool fully (resulting in the AI black box described by our expert librarians).

### 6.7 Exhaustion: Re-Search and Update

The final search should be conducted multiple times over the course of a few days to ensure that the resulting corpus size is stable. Further, new publications may appear during the process of selection, extraction and analysis. Although we disagree that any new publication automatically makes an SLR in HCI outdated, we nevertheless recommend that researchers at least consider doing a final update to their search prior to submission and publication. Depending on the review synthesis method, it may be possible to update the content somewhat easily. This approach, combined with the practice of reapplying the search strategy and noting any changes in the databases used, ensures the robustness and currency of the systematic review.

### 6.8 Limitations

The findings and the initial INSPIRE framework are grounded in the specialized knowledge of expert librarians. However, recruiting librarians who meet these criteria can be challenging. Our study was limited by a small sample size (n=8) due to the difficulty of finding librarians with both HCI experience and availability across time zones. We attempted to mitigate the effects of this by ensuring that the interviews were in-depth and allowing the interview conversations to continue freely and over the initially expected time (with the participant's explicit permission). The consistency of responses from these experts suggests potential saturation of the data, indicating that a larger sample may not have significantly altered the framework's core components. Additionally, due to the relatively small pool of librarians who fit these criteria, we are unable to disclose their locations/affiliations to protect their anonymity. We also acknowledge that all librarians are from the North American continent. This limits the generalizability of the framework within sub-domains of HCI. Results may differ, particularly for researchers at institutions with different access to library resources or support. Future research could also explore rigorous benchmarking methodologies essential to evaluate the performance of AI-powered search and selection.

While the INSPIRE framework for search is based on interviews, further studies involving researchers applying the framework and evaluating its impact on search comprehensiveness and review quality would strengthen the evidence base for its usefulness. As this is an initial framework, we will validate its utility for enhancing search comprehensiveness and review quality through future studies involving HCI practitioners and librarians.

## 7 Conclusion

In conclusion, the field of HCI currently faces challenges in conducting systematic literature searches due to the lack of a formalized framework and the diverse range of research strategies employed within the field. The reliance on methods derived from medical literature searches may not fully capture the nuances of HCI research or its common databases. In this context, we report on interviews with experts in accessing academic databases in the form of academic librarians and information specialists. Our analysis focuses on their recommendations for best practice, observations from conducting reviews in HCI and their views on computational tools in supporting the review search process. Based on this, we developed the INSPIRE framework as a tool to structure and support the search in literature reviews. Covering the stages of Inquiry, Navigation, Search, Point of Saturation, Involvement, Reporting and Exhaustion, the framework can guide HCI researchers through a comprehensive and rigorous literature review search using both divergent and convergent searching on keywords and keystrings.

## Acknowledgments

We would like to clarify that no automated computational tools, including AI, were used in the research process other than Grammarly, which we used to correct grammatical errors and improve writing style. This work was made possible by the NSERC Discovery Grant and the Canada Foundation for Innovation John R. Evans Leaders Fund.

We thank the Games Institute members at the University of Waterloo for all their support. We want to extend our heartfelt thanks to all who called the librarians and provided tremendous help! Lastly, we thank Tim Ireland, the liaison librarian at the University of Waterloo, for all their help in connecting us with the right people!

## Data Availability Statement

The supplementary data will also be hosted on our institutional website and the Open Science Framework: https://osf.io/39zyp/.

## References

Acar, O. *Pair Framework Guidance*. King's College London. https://www.kcl.ac.uk/about/strategy/learning-and-teaching/ai-guidance/pair-framework-guidance.

Allot, A., Lee, K., Chen, Q., Luo, L. and Lu, Z. (2021) LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res.*, **49**, W352–W358. https://doi.org/10.1093/nar/gkab326.

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. and Zayed, T. (2023) Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. *Systems*, **11**, 351. https://doi.org/10.3390/systems11070351.

Allen, B. (2000) Individual differences and the conundrums of user-centered design: two experiments. *J. Am. Soc. Inf. Sci.*, **51**, 508–520. https://doi.org/10.1002/(SICI)1097-4571(2000)51:6\((\)508::AID-ASI3\()\)3.0.CO;2-Q.

Armijo-Olivo, S., Craig, R. and Campbell, S. (2020) Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. *Res. Synth. Methods*, **11**, 484–493. https://doi.org/10.1002/jrsm.1398.

Atkinson, L. and Cipriani, A. (2018) How to carry out a literature search for a systematic review: a practical guide. *BJPsych Adv.*, **24**, 74–82. https://doi.org/10.1192/bja.2017.3.

Athukorala, K., Głowacka, D., Jacucci, G., Oulasvirta, A. and Vreeken, J. (2016) Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *J. Assoc. Inf. Sci. Technol.*, **67**, 2635–2651. https://doi.org/10.1002/asi.23617.

Bailey, J., Zhang, C., Budgen, D., Turner, M. and Charters, S. (2007) Search engine overlaps: do they agree or disagree? *Second*

*International Workshop On Realising Evidence-Based Software Engineering (REBSE'07)*, 2–2. https://doi.org/10.1109/REBSE.2007.4.

Banathy, B. (1996) *Designing Social Systems in a Changing World*. Springer Science and Business Media.

Bannigan, K. and Watson, R. (2009) Reliability and validity in a nutshell. *J. Clin. Nurs.*, **18**, 3237–3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x.

Beller, E.*et al.* (2018) Principles of the international collaboration for the automation of systematic Reviews (ICASR). *Syst. Rev.*, **7**, 1–7. https://doi.org/10.1186/s13643-018-0740-7.

Benzies, K., Premji, S., Hayden, K. and Serrett, K. (2006) State-of-the-evidence reviews: advantages and challenges of including Grey literature. *Worldviews Evid.-Based Nurs.*, **3**, 55–61. https://doi.org/10.1111/j.1741-6787.2006.00051.x.

Blaizot, A., Veettil, S., Saidoung, P., Moreno-Garcia, C., Wiratunga, N., Aceves-Martins, M., Lai, N. and Chaiyakunapruk, N. (2022) Using artificial intelligence methods for systematic review in health sciences: a systematic review. *Res. Synth. Methods*, **13**, 353–362. https://doi.org/10.1002/jrsm.1553.

Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R. & Cleven, A. Reconstructing the giant: on the importance of rigour in documenting the literature search process. (2009), https://aisel.aisnet.org/ecis2009/161.

Bommasani, R., Hudson, D., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M., Bohg, J., Bosselut, A. and Brunskill, E. (2021) Others on the opportunities and risks of foundation models. ArXiv Preprint ArXiv:2108.07258.

Boucher, V. (1997) *Interlibrary Loan Practices Handbook*. American Library Association.

Braun, V. and Clarke, V. (2021) Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Couns. Psychother. Res.*, **21**, 37–47. https://doi.org/10.1002/capr.12360.

Brown, S., Hutton, B., Clifford, T., Coyle, D., Grima, D., Wells, G. and Cameron, C. (2014) A Microsoft-excel-based tool for running and critically appraising network meta-analyses—an overview and application of NetMetaXL. *Syst. Rev.*, **3**, 1–11. https://doi.org/10.1186/2046-4053-3-110.

Reviews, C. and Dissemination Systematic reviews (2009) CRD's guidance for undertaking reviews in health care. *Centre for Reviews*. https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf.

Chai, K., Lines, R., Gucciardi, D. and Ng, L. (2021) Research screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Syst. Rev.*, **10**, 93–13. https://doi.org/10.1186/s13643-021-01635-3.

Chapman, A., Morgan, L. and Gartlehner, G. (2010) Semi-automating the manual literature search for systematic reviews increases efficiency. *Health Inf. Libr. J.*, **27**, 22–27. https://doi.org/10.1111/j.1471-1842.2009.00865.x.

Chetwynd, E. (2022) Critical analysis of reliability and validity in literature reviews. *J. Hum. Lact.*, **38**, 392–396. https://doi.org/10.1177/08903344221100201.

Choong, M., Galgani, F., Dunn, A. and Tsafnat, G. (2014) Automatic evidence retrieval for systematic reviews. *J. Med. Internet Res.*, **16**, e223. https://doi.org/10.2196/jmir.3369.

Cierco Jimenez, R., Lee, T., Rosillo, N., Cordova, R., Cree, I., Gonzalez, A. and Indave Ruiz, B. (2022) Machine learning computational tools to assist the performance of systematic reviews: a mapping review. *BMC Med. Res. Methodol.*, **22**, 322. https://doi.org/10.1186/s12874-022-01805-4.

Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P. and Scott, A. (2020) A full systematic review was completed in 2 weeks using automation tools: a case study. *J. Clin. Epidemiol.*, **121**, 81–90. https://doi.org/10.1016/j.jclinepi.2020.01.008.

Cleo, G., Scott, A., Islam, F., Julien, B. and Beller, E. (2019) Usability and acceptability of four systematic review automation software packages: a mixed method design. *Syst. Rev.*, **8**, 145–145. https://doi.org/10.1186/s13643-019-1069-6.

Cooke, A., Smith, D. and Booth, A. (2012) Beyond PICO: the SPIDER tool for qualitative evidence synthesis. *Qual. Health Res.*, **22**, 1435–1443. https://doi.org/10.1177/1049732312452938.

Cooper, C., Booth, A., Varley-Campbell, J., Britten, N. and Garside, R. (2018) Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. *BMC Med. Res. Methodol.*, **18**, 85–14. https://doi.org/10.1186/s12874-018-0545-3.

Cooper, H., Hedges, L. and Valentine, J. (2019) *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.

Corrall, S. (2012) Roles and responsibilities: libraries, librarians and data. *Manag. Res. Data*, 105–133. https://doi.org/10.29085/9781856048910.007.

Desmeules, R., Campbell, S. and Dorgan, M. (2016) Acknowledging librarians' contributions to systematic review searching. *J. Canad. Health Lib. Assoc.*, **37**. https://doi.org/10.5596/c16-014.

Reis, A., Oliveira, A., Fritsch, C., Zouch, J., Ferreira, P. and Polese, J. (2023) Usefulness of machine learning softwares to screen titles of systematic reviews: a methodological study. *Syst. Rev.*, **12**, 68. https://doi.org/10.1186/s13643-023-02231-3.

Echtler, F. and Häußler, M. (2018) Open source, open science, and the replication crisis in HCI. *Extended Abstracts Of The 2018 CHI Conference On Human Factors In Computing Systems*, 1–8. https://doi.org/10.1145/3170427.3188395.

Elliott, J., Turner, T., Clavisi, O., Thomas, J., Higgins, J., Mavergames, C. and Gruen, R. (2014) Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Med.*, **11**, e1001603. https://doi.org/10.1371/journal.pmed.1001603.

Feger, S., Dallmeier-Tiessen, S., Woźniak, P. & Schmidt, A. The Role of hci in reproducible science: understanding, supporting and motivating core practices. *Extended Abstracts Of The 2019 CHI Conference On Human Factors In Computing Systems*. pp. 1–6 (2019), https://doi.org/10.1145/3290607.3312905

Felizardo, K. and Carver, J. (2020) Automating systematic literature review. *Contemporary Empirical Methods In Software Engineering*, 327–355. https://doi.org/10.1007/978-3-030-32489-6_12.

Finfgeld-Connett, D. and Johnson, E. (2013) Literature search strategies for conducting knowledge-building and theory-generating qualitative systematic reviews. *J. Adv. Nurs.*, **69**, 194–204. https://doi.org/10.1111/j.1365-2648.2012.06037.x.

Friedman, L., Furberg, C., DeMets, D., Reboussin, D. and Granger, C. (2015) *Fundamentals of Clinical Trials*. Springer.

Gartlehner, G., Wagner, G., Lux, L., Affengruber, L., Dobrescu, A., Kaminski-Hartenthaler, A. and Viswanathan, M. (2019) Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. *Syst. Rev.*, **8**, 277–210. https://doi.org/10.1186/s13643-019-1221-3.

Gates, A., Vandermeer, B. and Hartling, L. (2018) Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the RobotReviewer machine learning tool. *J. Clin. Epidemiol.*, **96**, 54–62. https://doi.org/10.1016/j.jclinepi.2017.12.015.

Gates, A., Gates, M., Sebastianski, M., Guitard, S., Elliott, S. and Hartling, L. (2020) The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and

rapid reviews. *BMC Med. Res. Methodol.*, **20**, 139–139. https://doi.org/10.1186/s12874-020-01031-w.

Gross, T. & Taylor, A. (2005) What have we got to lose? The effect of controlled vocabulary on keyword searching results. *Coll. Res. Libr.*, https://repository.stcloudstate.edu/lrs_facpubs/13/, **66**, 212, 230, https://doi.org/10.5860/crl.66.3.212.

Gundersen, O. E. and Kjensmo, S. (2018) State of the art: reproducibility in artificial intelligence. *Proc. AAAI Conf. Artif. Intell.*, **32**. https://doi.org/10.1609/aaai.v32i1.11503.

Gusenbauer, M. (2022) Search where you will find most: comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics*, **127**, 2683–2745. https://doi.org/10.1007/s11192-022-04289-7.

Gusenbauer, M. and Haddaway, N. (2021) What every researcher should know about searching–clarified concepts, search advice, and an agenda to improve finding in academia. *Res. Synth. Methods*, **12**, 136–147. https://doi.org/10.1002/jrsm.1457.

Hamel, C., Michaud, A., Thuku, M., Skidmore, B., Stevens, A., Nussbaumer-Streit, B. and Garritty, C. (2021) Defining rapid reviews: a systematic scoping review and thematic analysis of definitions and defining characteristics of rapid reviews. *J. Clin. Epidemiol.*, **129**, 74–85. https://doi.org/10.1016/j.jclinepi.2020.09.041.

Hancock, J., Naaman, M. and Levy, K. (2020) AI-mediated communication: definition, research agenda, and ethical considerations. *J. Comput.-Mediat. Commun.*, **25**, 89–100. https://doi.org/10.1093/jcmc/zmz022.

Herrmannova, D., Patton, R., Knoth, P. and Stahl, C. (2018) Do citations and readership identify seminal publications? *Scientometrics*, **115**, 239–262. https://doi.org/10.1007/s11192-018-2669-y.

Hirt, J., Nordhausen, T., Appenzeller-Herzog, C. and Ewald, H. (2021) Using citation tracking for systematic literature searching—study protocol for a scoping review of methodological studies and a Delphi study [version 3; peer review: 2 approved]. *F1000Research*, **9**, 1386. https://doi.org/10.12688/f1000research.27337.3.

Hornbæk, K., Sander, S., Bargas-Avila, J. and Grue Simonsen, J. (2014) Is once enough? On the extent and content of replications in human–computer interaction. *Proceedings Of The SIGCHI Conference On Human Factors In Computing Systems*, 3523–3532. https://doi.org/10.1145/2556288.25570040.

Ibrahim, Z., Karimi, P., Martin-Hammond, A., Harrington, C. and Siek, K. (2024) What do we do? Lessons learned from conducting systematic reviews to improve HCI dissemination. *Extended Abstracts Of The CHI Conference On Human Factors In Computing Systems*, 1–8. https://doi.org/10.1145/3613905.3637117.

Jalali, S. and Wohlin, C. (2012) Systematic literature studies: database searches vs. backward snowballing. *Proceedings Of The ACM-IEEE International Symposium On Empirical Software Engineering And Measurement*, 29–38. https://doi.org/10.1145/2372251.2372257.

Janssens, A., Gwinn, M., Brockman, J., Powell, K. and Goodman, M. (2020) Novel citation-based search method for scientific literature: a validation study. *BMC Med. Res. Methodol.*, **20**, 1–11. https://doi.org/10.1186/s12874-020-0907-5.

Jordan, Z., Lockwood, C., Munn, Z. and Aromataris, E. (2019) The updated Joanna Briggs institute model of evidence-based healthcare. *JBI Evid. Implement.*, **17**, 58–71. https://doi.org/10.1097/XEB.0000000000000155.

Khalil, H., Ameen, D. and Zarnegar, A. (2022) Tools to support the automation of systematic reviews: a scoping review. *J. Clin. Epidemiol.*, **144**, 22. https://doi.org/10.1016/j.jclinepi.2021.12.005.

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A. and Hadfield, K. (2024) Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res. Synth. Methods*, **15**, 616–626. https://doi.org/10.1002/jrsm.1715.

Kraus, S., Breier, M., Lim, W., Dabić, M., Kumar, S., Kanbach, D., Mukherjee, D., Corvello, V., Piñeiro-Chousa, J. and Liguori, E. (2022) Others literature reviews as independent studies: guidelines for academic practice. *Rev. Manag. Sci.*, **16**, 2577–2595. https://doi.org/10.1007/s11846-022-00588-8.

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A., Hammerstrøm, K. and Sathe, N. (2016) Searching for studies: a guide to information retrieval for Campbell. *Campbell Syst. Rev.*, **13**, 1–73. https://doi.org/10.4073/cmg.2016.1.

Lefebvre, C., Manheimer, E. and Glanville, J. (2008) Searching for studies. *Cochrane Handbook For Systematic Reviews Of Interventions: Cochrane Book Series*, 95–150. https://doi.org/10.1002/9780470712184.ch6.

Liang, S., He, D., Wu, D. and Hu, H. (2020) Challenges and opportunities of ACM digital library: a preliminary survey on different users. In *Sustainable Digital Communities: 15th International Conference, IConference 2020, Boras, Sweden, March 23–26, 2020, Proceedings 15*, pp. 278–287.

Lo, L. (2023) The CLEAR path: a framework for enhancing information literacy through prompt engineering. *J. Acad. Librariansh.*, **49**, 102720. https://doi.org/10.1016/j.acalib.2023.102720.

Lund, B., Wang, T., Mannuru, N., Nie, B., Shimray, S. and Wang, Z. (2023) ChatGPT and a new academic reality: artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J. Assoc. Inf. Sci. Technol.*, **74**, 570–581. https://doi.org/10.1002/asi.24750.

Mahmoudi, H., Chang, D., Lee, H., Ghaffarzadegan, N. and Jalali, M. (2024) A critical assessment of large language models for systematic reviews: utilizing ChatGPT for complex data extraction. Available At SSRN 4797024. https://doi.org/10.2139/ssrn.4797024.

Marchionini, G. (2006) Exploratory search: from finding to understanding. *Commun. ACM*, **49**, 41–46. https://doi.org/10.1145/1121949.1121979.

Marshall, I. and Wallace, B. (2019) Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst. Rev.*, **8**, 163–110. https://doi.org/10.1186/s13643-019-1074-9.

Martinez Garcia, L., McFarlane, E., Barnes, S., Sanabria, A., Alonso–Coello, P. and Alderson, P. (2014) Updated recommendations: an assessment of NICE clinical guidelines. *Implement. Sci.*, **9**, 1–9. https://doi.org/10.1186/1748-5908-9-72.

McGowan, J., Sampson, M., Salzwedel, D., Cogo, E., Foerster, V. and Lefebvre, C. (2016) PRESS peer review of electronic search strategies: 2015 guideline statement. *J. Clin. Epidemiol.*, **75**, 40–46. https://doi.org/10.1016/j.jclinepi.2016.01.021.

Meert, D., Torabi, N. and Costella, J. (2016) Impact of librarians on reporting of the literature searching component of pediatric systematic reviews. *J. Med. Lib. Assoc.*, **104**, 267–277. https://doi.org/10.3163/1536-5050.104.4.004.

Monarch, R. (2021) *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. Simon.

Moher, D., Cook, D., Eastwood, S., Olkin, I., Rennie, D. & Stroup, D. (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet*, **354**, 1896–1900, https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(99)04149-5/fulltext, https://doi.org/10.1016/S0140-6736(99)04149-5

Munn, Z.*et al.* (2019) The Joanna Briggs institute system for the unified management, assessment and review of information (JBI

SUMARI). *JBI Evid. Implement.*, **17**, 36–43. https://doi.org/10.1097/XEB.0000000000000152.

Naderifar, M., Goli, H. and Ghaljaie, F. (2017) Snowball sampling: a purposeful method of sampling in qualitative research. *Strides Dev. Med. Educ.*, **14**. https://doi.org/10.5812/sdme.67670.

Neimann Rasmussen, L. and Montgomery, P. (2018) The prevalence of and factors associated with inclusion of non-English language studies in Campbell systematic reviews: a survey and meta-epidemiological study. *Syst. Rev.*, **7**, 129–112. https://doi.org/10.1186/s13643-018-0786-6.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. and Ananiadou, S. (2015) Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst. Rev.*, **4**, 1–22. https://doi.org/10.1186/2046-4053-4-5.

Okoli, C. (2015) A guide to conducting a standalone systematic literature review. *Commun. Assoc. Inf. Syst.*, **37**. https://doi.org/10.17705/1CAIS.03743.

Ouzzani, M., Hammady, H., Fedorowicz, Z. and Elmagarmid, A. (2016) Rayyan—a web and mobile app for systematic reviews. *Syst. Rev.*, **5**, 210–210. https://doi.org/10.1186/s13643-016-0384-4.

Paez, A. (2017) Gray literature: an important resource in systematic reviews. *J. Evid.-Based Med.*, **10**, 233–240. https://doi.org/10.1111/jebm.12266.

Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., Shamseer, L., Tetzlaff, J. and Moher, D. (2021a) Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *J. Clin. Epidemiol.*, **134**, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003.

Page, M. J.*et al.* (2021b) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, **372**, n71. https://doi.org/10.1136/bmj.n71.

Petticrew, M. and Roberts, H. (2008) *Systematic Reviews in the Social Sciences: A Practical Guide*. John Wiley and Sons.

Pham, B.*et al.* (2021) A semi-automated workflow. *Syst. Rev.*, **10**, 156. https://doi.org/10.1186/s13643-021-01700-x.

Przybyła, P., Brockmeier, A., Kontonatsios, G., Le Pogam, M., McNaught, J., Elm, E., Nolan, K. and Ananiadou, S. (2018) Prioritising references for systematic reviews with RobotAnalyst: a user study. *Res. Synth. Methods*, **9**, 470–488. https://doi.org/10.1002/jrsm.1311.

Rethlefsen, M. L.*et al.* (2021) PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Syst. Rev.*, **10**, 39. https://doi.org/10.1186/s13643-020-01542-z.

Romanelli, J., Gonçalves, M., Abreu Pestana, L., Soares, J., Boschi, R. and Andrade, D. (2021) Four challenges when conducting bibliometric reviews and how to deal with them. *Environ. Sci. Pollut. Res.*, **28**, 60448–60458. https://doi.org/10.1007/s11356-021-16420-x.

Rogers, K. The shiny scary future of automated research synthesis in HCI. *Accepted Paper at the CHI '24 Workshop on LLMs as Research Tools: Applications and Evaluations in HCI Data Work*. https://osf.io/txwzy.

Rogers, K. and Seaborn, K. (2023) The systematic review-lution: a manifesto to promote rigour and inclusivity in research synthesis. *Extended Abstracts Of The 2023 CHI Conference On Human Factors In Computing Systems*. https://doi.org/10.1145/3544549.3582733.

Rogers, K., Hirzle, T., Karaosmanoglu, S., Palomino, P., Durmanova, E., Isotani, S. and Nacke, L. (2024) An umbrella review of reporting quality in CHI systematic reviews: guiding questions and best practices for HCI. *ACM Trans. Comput.-Hum. Interact.*, **31**, 1–55. https://doi.org/10.1145/3685266.

Ros, R., Bjarnason, E. and Runeson, P. (2017) A machine learning approach for semi-automated search and selection in literature studies. *Proceedings Of The 21st International Conference On Evaluation And Assessment In Software Engineering*, 118–127. https://doi.org/10.1145/3084226.3084243.

Rubin, R. and Rubin, R. (2020) *Foundations of Library and Information Science*. American Library Association.

Safdar, N., Banja, J. and Meltzer, C. (2020) Ethical considerations in artificial intelligence. *Eur. J. Radiol.*, **122**, 108768. https://doi.org/10.1016/j.ejrad.2019.108768.

Samek, W. and Müller, K.-R. (2019) Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 5–22. Springer International Publishing, Cham.

Schardt, C., Adams, M., Owens, T., Keitz, S. and Fontelo, P. (2007) Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med. Inform. Decis. Mak.*, **7**, 1–6. https://doi.org/10.1186/1472-6947-7-16.

Shojania, K., Sampson, M., Ansari, M., Ji, J., Doucette, S. and Moher, D. (2007) How quickly do systematic reviews go out of date? A survival analysis. *Ann. Intern. Med.*, **147**, 224–233. https://doi.org/10.7326/0003-4819-147-4-200708210-00179.

Smucker, M. and Allan, J. (2006) Find-similar: similarity browsing as a search tool. *Proceedings Of The 29th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval*, 461–468. https://doi.org/10.1145/1148170.1148250.

Spencer, A. and Eldredge, J. (2018) Roles for librarians in systematic reviews: a scoping review. *J. Med. Lib. Assoc.*, **106**, 46–56. https://doi.org/10.5195/jmla.2018.82.

Stahl, B., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Shaelou, S., Patel, A., Ryan, M. and Wright, D. (2021) Artificial intelligence for human flourishing–beyond principles for machine learning. *J. Bus. Res.*, **124**, 374–388. https://doi.org/10.1016/j.jbusres.2020.11.030.

Stefanidi, E., Bentvelzen, M., Woźniak, P., Kosch, T., Woźniak, M., Mildner, T., Schneegass, S., Müller, H. and Niess, J. (2023) Literature reviews in HCI: a review of reviews. *Proceedings Of The 2023 CHI Conference On Human Factors In Computing Systems*, 1–24. https://doi.org/10.1145/3544548.3581332.

Tsafnat, G., Dunn, A., Glasziou, P. and Coiera, E. (2013) The automation of systematic reviews. *BMJ*, **346**, f139–f139. https://doi.org/10.1136/bmj.f139.

Tsafnat, G., Glasziou, P., Choong, M., Dunn, A., Galgani, F. and Coiera, E. (2014) Systematic review automation technologies. *Syst. Rev.*, **3**, 1–15. http://www.systematicreviewsjournal.com/content/3/1/74.

Uman, L. (2011) Systematic reviews and meta-analyses. *J. Can. Acad. Child Adolesc. Psychiatry*, **20**, 57. https://pmc.ncbi.nlm.nih.gov/articles/PMC3024725/.

Valizadeh, A., Moassefi, M., Nakhostin-Ansari, A., Hosseini Asl, S., Saghab Torbati, M., Aghajani, R., Maleki Ghorbani, Z. & Faghani, S. (2022) Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. *BMC Med. Res. Methodol.* **22**, 160, https://doi.org/10.1186/s12874-022-01631-8

Van den Brand, S. A. G. E. and van de Schoot, R. A systematic review on studies evaluating the performance of active learning compared to human reading for systematic review data. https://doi.org/10.17605/OSF.IO/T9HGM.

Van Dinter, R., Tekinerdogan, B. and Catal, C. (2021a) Automation of systematic literature reviews: a systematic literature review. *Inf. Softw. Technol.*, **136**, 106589. https://doi.org/10.1016/j.infsof.2021.106589.

Van Dinter, R., Catal, C. and Tekinerdogan, B. (2021b) A multi-channel convolutional neural network approach to automate the citation screening process. *Appl. Soft Comput.*, **112**, 107765. https://doi.org/10.1016/j.asoc.2021.107765.

Van Dijk, S., Brusse-Keizer, M., Bucsán, C., Palen, J., Doggen, C. and Lenferink, A. (2023) Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ Open*, **13**, e072254. https://doi.org/10.1136/bmjopen-2023-072254.

Schoot, R.*et al.* (2021) An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Mach. Intell.*, **3**, 125–133. https://doi.org/10.1038/s42256-020-00287-7.

Wagner, G., Lukyanenko, R. and Paré, G. (2022) Artificial intelligence and the conduct of literature reviews. *J. Inf. Technol.*, **37**, 209–226. https://doi.org/10.1177/02683962211048201.

Weber, H., Becker, D. and Hillmert, S. (2019) Information-seeking behaviour and academic success in higher education: which search strategies matter for grade differences among university students and how does this relevance differ by field of study? *High. Educ.*, **77**, 657–678. https://doi.org/10.1007/s10734-018-0296-4.

White, R. and Roth, R. (2009) *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan and Claypool Publishers.

Wohlin, C., Kalinowski, M., Felizardo, K. and Mendes, E. (2022) Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Inf. Softw. Technol.*, **147**, 106908. https://doi.org/10.1016/j.infsof.2022.106908.

Wu, C., Chakravorti, T., Carroll, J. M. and Rajtmajer, S. (2024) Integrating measures of replicability into scholarly search: challenges and opportunities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, pp. 1–18. Association for Computing Machinery, New York, NY, USA, Article 18.

Zhang, Y., Liang, S., Feng, Y., Wang, Q., Sun, F., Chen, S., Yang, Y., He, X., Zhu, H. and Pan, H. (2022) Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. *Syst. Rev.*, **11**, 11–17. https://doi.org/10.1186/s13643-021-01881-5.

Zheng, C., Yuan, K., Guo, B., Hadi Mogavi, R., Peng, Z., Ma, S. and Ma, X. (2024) Charting the future of AI in project-based learning: a Co-design exploration with students. *Proceedings Of The CHI Conference On Human Factors In Computing Systems*, 1–19. https://doi.org/10.1145/3613904.3642807.

Zwakman, M., Verberne, L., Kars, M., Hooft, L., Delden, J. and Spijker, R. (2018) Introducing PALETTE: an iterative method for conducting a literature search for a review in palliative care. *BMC Palliat. Care*, **17**, 82–89. https://doi.org/10.1186/s12904-018-0335-z.

## APPENDIX A

In this section, we provide additional explanations on how the analysis is conducted to align with transparency principles.

The supplementary data will also be hosted on our institutional website and the Open Science Framework: https://osf.io/39zyp/.

In Fig. A1, we present the initial codes used by the first author in the initial raw line-by-line coding in Dovetail, prior to any refinement. We note that these codes and very early-stage categories do not represent a codebook or themes.

We elaborate on an example to clarify the process applied, using the second-order code 'keywords authors provided' as a starting point—see Fig. A2. The librarians reported authors' tendency to bring relevant keywords to the process of undertaking a formalized literature review, so we applied this tag 'keywords authors provided'. These keywords are grounded in the authors' subject knowledge, including subject headings and keywords supplied by the authors of relevant papers. We used a combination of first-order and second-order codes. Our above example is a second-order code that was developed based on the first-order codes of: 'subject headings are important', 'author-supplied keywords are useful' and 'subject searches provided', see Fig. A3.

Collectively, these codes highlight the importance of subject headers and keywords in the search and selection segments, emphasizing that these elements should be informed by the research team's inherent domain knowledge. This connection is significant as it relates to second-order ('hierarchical') coding, which integrates and organizes these ideas in a structured manner, ensuring that the initial coding is mapped to broader themes that guide and refine the research process for a more targeted and effective literature review.

The second-order codes were still closely connected to the participant responses, which in turn were structured based on the questions asked in the interview (see supplementary data). The particular responses tagged by these codes came from the probing question: '*Can you share a bit about your role and experience as a librarian, especially in the context of literature searching?*'. We acknowledge that this process is mainly semantic and inductive tags; because these librarians are experts, we aimed to keep as close to their meaning as possible and refrain from more latent coding approaches.

Further on, we created a codebook from these initial first- and second-order codes. In this step, this specific second-order code was subsumed into a new code of 'subject headings'. Later, in Miro, this code became part of the theme 'Hooking The Right Key Words', as developed by the first author into more narratively-shaped themes in an iterative process.

At first, the theme included the word 'Author' (i.e. Hooking Keywords Author-Supplied). However, the phrasing was not clear, and this did not fully reflect our data because the keyword selection is shaped not only by the authors on the research team, who provide domain knowledge, but also by librarians who contribute as methodologist experts. Yet, we found that sometimes the librarians are not included as authors on the formalized literature review (sometimes because of internal or funding-related politics or guidelines; we do not disclose further details as this could reveal institutional policies and break confidentiality).

## Positionality Statement

As a multidisciplinary research team with diverse academic and professional backgrounds, we recognize that our experiences shape our approach to exploring best practices in Human–Computer Interaction (HCI) for conducting formalized literature reviews (including SLRs) within HCI. Our goal is to enhance methodological rigour in HCI by addressing challenges in the search and selection stages of SRs. Our expertise spans game user research, knowledge synthesis and user engagement. However, we acknowledge that our positions—primarily shaped by our institutional affiliations in North America and Europe—influence our perspectives.

## Interview Questions

It is important to note that not all questions were explicitly asked or asked in the same order. Instead, if the librarian provided strategies on search and selection (and relating to our RQ) unprompted or in response to another question, the researchers would go off script and probe promptly (not in chronological order by any means). Questions already covered in previous responses were not re-asked.

The question pool was semi-structured and focused on asking about their Experience As A Librarian, Understanding The RQ,

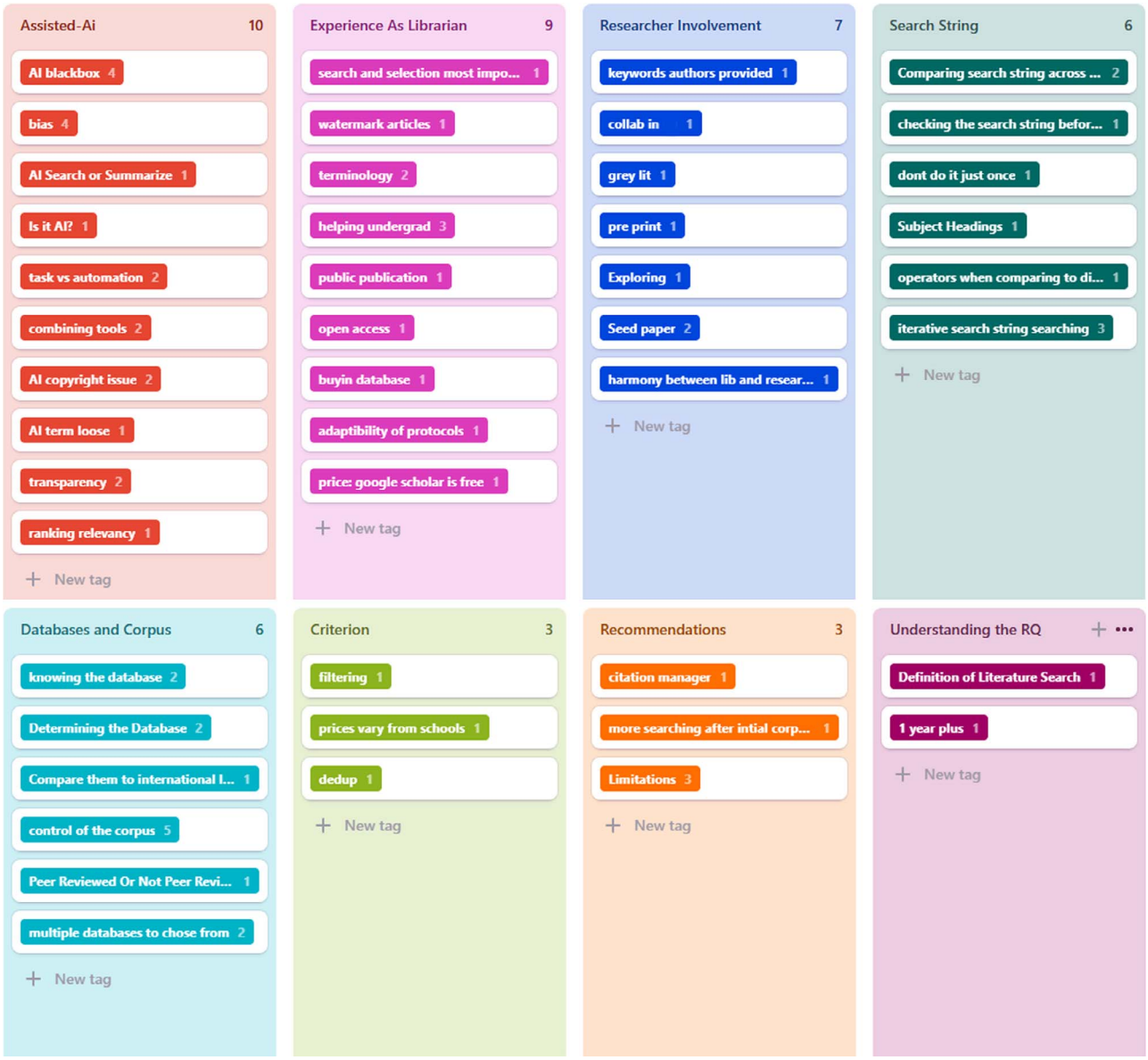


**Fig. A1.** This figure illustrates the first batch of initial tags based on the data of three randomly-chosen participants.

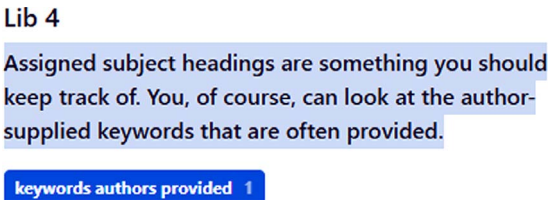


**Fig. A2.** This figure illustrates an example tag of line-by-line initial tagging.

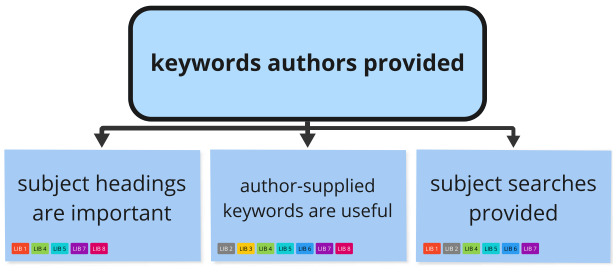


**Fig. A3.** This figure illustrates an example of the flat and hierarchical codes.

Researcher Involvement, Databases and Corpus, Search String, Criterion, Recommendations and Assisted Tools. The question pool can be found on the Open Science Framework.

## INSPIRE Framework

We include a graphical figure of the initial INSPIRE framework (a high-quality version of this figure will be available in our institutional archive for better resolution and accessibility) see Fig. A4.

Our framework differs from the PRISMA-S extension (Rethlefsen, 2021) and the PALETTE method (Zwakman *et al.*, 2018) in key ways. While PRISMA-S focuses on improving transparency and reproducibility in the search strategy, our framework allows for a more iterative, flexible process where stages like Inquiry and Reporting may overlap and be revisited. The PALETTE method offers a structured approach for systematic reviews, but our INSPIRE

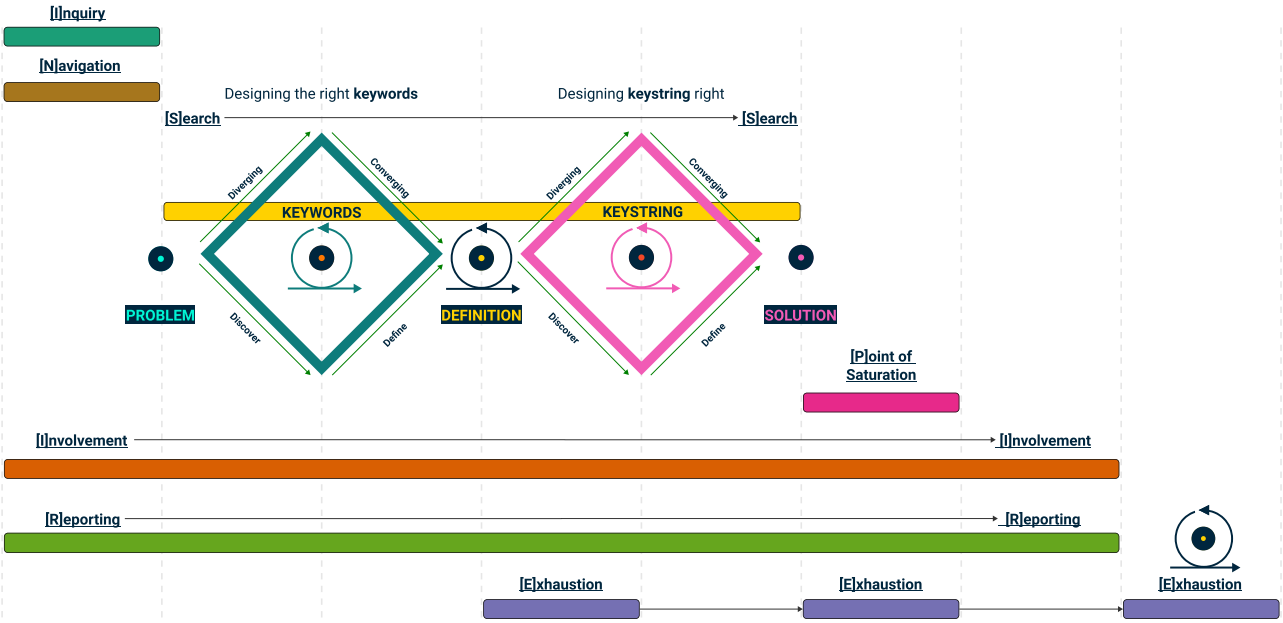


**Fig. A4.** This figure presents the initial INSPIRE framework, developed based on insights from interviews with expert librarians, to streamline the search and selection stages in formalized literature reviews. The framework comprises seven interconnected segments: Inquiry, Navigation, Search, Point of Saturation, Involvement, Reporting and Exhaustion. It consists of seven interconnected segments: Inquiry, Navigation, Search, Point of Saturation, Involvement, Reporting and Exhaustion. These segments are not intended to follow a strict sequential order but rather reflect an iterative and parallel approach where stages may overlap, pause and be revisited as needed.

framework emphasizes real-time involvement and dynamic adjustments, particularly in stages like Point of Saturation and Exhaustion.

## APPENDIX B - Checklist

In this section, we provide a supplementary checklist for researchers to consider asking themselves for each segment of the initial INSPIRE framework.

### Inspire

To truly inspire their literature review, researchers should ask themselves these questions in this initial segment:

- Do these goals align with the intended research method (e.g. scoping review, full systematic review)?
- Does the team possess relevant background knowledge or expertise in the HCI field?
- (When possible) Has the team considered an information specialist (librarian) with experience in knowledge synthesis and information retrieval?
- Are clear and concise research goals formulated to guide the review?
- Does at least one team member have familiarity with the specific HCI topic to develop targeted search strings and terminology?
- Has the team conducted a preliminary search to identify existing or ongoing reviews in the chosen HCI topic?

### Navigation

To ensure a thorough navigation stage, here are crucial questions researchers should ask themselves:

- Do you have a set of seed paper(s) or foundational work(s) relevant to your research topic?
- Does the team have access to relevant databases through institutional subscriptions or interlibrary loan (ILL) networks?
- What databases or publication platforms were the seed paper(s) published in?
- What are the typical types of publications (journals, conferences, etc.) that house research in this domain?
- Have you identified a list of relevant academic databases based on your research domain and seed paper(s)?
- Have you reviewed the search functionalities and coverage of each chosen database?
- Have you considered the possibility of cross-referencing between databases to ensure you are capturing the seed paper(s)?
- Have you reviewed the citation history (forward/backwards/both) of your seed paper(s) to identify earlier foundational works highly cited within your field?

### Search

In summary, the questions one should ask themselves in searching approaches are:

- Have you analyzed the content and abstracts of your seed papers to extract key terms and indexing vocabulary used by the authors?
- Have you incorporated convergence searching to refine your search terms and focus on highly relevant studies?
- Have you also employed divergent searching to broaden your search and explore new avenues for relevant research by identifying related concepts, synonyms and alternative keywords based on your initial findings?
- As you identify new studies, are you iteratively refining your search terms and strategies to ensure a comprehensive search?
- Have you considered filtering by subdomains (subject

headings) from relevant databases to refine further and categorize your search?
- Are you planning to utilize citation chasing to explore references from identified studies and potentially discover foundational works you might have missed?

## Point of Saturation

Some questions to think about during the point of saturation stage are:

- Does the database selection support the comprehensiveness of coverage for the research topic?
- When reducing the scope of years searched, is a pivotal point (e.g. the publication year of a key seed article) used to ensure recent and relevant studies are captured?
- Does the search string consider the specific context and technical jargon of the research question?
- Does the search strategy account for inconsistencies in terminology and character limitations within different databases?
- Has the researcher considered limitations such as Boolean operator functionality and wildcard support across different databases to ensure consistency in search results?

## Involvement

Some questions to think about during the involvement stage:

- (When possible) Have you considered with a librarian specializing in HCI and information science to identify relevant databases, subject headings and terminology?
- Are you aware of potential issues with watermarked articles accessed through university VPNs when using AI tools?
- If necessary, have you identified resources or training opportunities to improve your team's information literacy skills (e.g. the terminologies)?
- Are you aware of the limitations of AI tools regarding transparency and replicability?

## Reporting

Some questions to think about during the reporting stages are:

- Does the report clearly detail the search strategies used in the literature review?
- Is the rationale behind the choice of key search terms explained in detail?
- Does the report address maintaining the rigour of the review process through transparent reporting?
- If computational tools (including AI) are used, does the report demonstrate a clear understanding of their functionality (algorithms, data sources, ranking criteria) and limitations?

## Exhaustion

Final questions to ensure the search and selection process is exhaustive:

- Have you set up alerts on relevant academic databases to receive notifications about new publications related to your topic?
- Have you scheduled regular revisits to your initial search terms to identify potentially significant new findings?