SUPPORTING SCHOLARLY RESEARCH IDEATION THROUGH WEB SEMANTICS

A Dissertation

by

YIN QU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Andruid Kerne
Committee Members,	Steven M. Smith
	James Caverlee
	Thomas Ioerger
Head of Department,	Dilma da Silva

August 2019

Major Subject: Computer Science

ABSTRACT

We develop new methods and technologies for supporting *scholarly research ideation*, the tasks in which researchers develop new ideas for their work, through *web semantics*, computational representations of information found on the web, capturing meaning involving people's experiences of things of interest. To do so, we first conducted a qualitative study with established researchers on their practices, using sensitizing concepts from information science, creative cognition, and art as a basis for framing and deriving findings. We found that participants engage in and combine a wide range of activities, including *citation chaining*, *exploratory browsing*, and *curation*, to achieve their goals of creative ideation. We derived a new, interdisciplinary model to depict their practices. Our study and findings address a gap in existing research: the creative nature of what researchers do has been insufficiently investigated. The model is expected to guide future investigations.

We then use in-context presentations of dynamically extracted semantic information to (1) address the issues of digression and disorientation, which arise in citation chaining and exploratory browsing, and (2) provide contextual information in researchers' prior work curation. The implemented interface, Metadata In-Context Explorer (MICE), maintains context while allowing new information to be brought into and integrated with the current context, reducing the needs for switching between documents and webpages. Study shows that MICE supports participants in their citation chaining processes, thus supports scholarly research ideation. MICE is implemented with BigSemantics, a metadata type system and runtime integrating data models, extraction rules, and presentation hints into *types*. BigSemantics operationalizes type-specific, dynamic extraction and rich presentation of semantic information (a.k.a. metadata) found on the web. The metadata type system, runtime, and MICE are expected to help build interfaces supporting dynamic exploratory search, browsing, and other creative tasks involving complex and interlinked semantics.

ACKNOWLEDGMENTS

I would like to thank my advisor, Andruid Kerne, for his support for and mentoring of my PhD study. Without you, this would not be possible. Your knowledge, your way of thinking, and your values all deeply influence me.

I would also like to thank my committee for providing feedback. Special thanks to Cathy Marshall, who served as a special appointment on my committee and gave me many valuable advices making this work better.

Thanks to Nic Lupfer, Rhema Linder, Ajit Jain, and Matthew Carrasco, who made major direct contributions to this work. Nic developed the MICE interface. Rhema helped with the statistics. Ajit and Matthew collaborated with me on conducting the study with researchers.

I also received immense help and precious friendship from other members in the lab: Andrew Webb, Bill Hamilton, Shengfeng Fei, Feiyu Yu, among many others. This dissertation also builds upon the work of Abhinav Mathur, Nabeel Shahzad, Sashikanth Damaraju, Kade Keith, Zach Brown, Tylor Tesch, and Tom White. I had a great time working and having fun with all of them.

Last but not least, thanks to my parents, and my wife, Yue, for their unconditional support and love. I feel so lucky and grateful to have you in my life.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Andruid Kerne, Professor Steven Smith (of the Department of Psychology), Professor James Caverlee, Professor Thomas Ioerger, and Professor Cathy Marshall.

As a team, members in our lab made indispensible contributions to this work. Ajit Jain and Matthew Carrasco contributed to the qualitative analysis in Chapter 3. Nic Lupfer developed the web-based MICE interface. Rhema Linder conducted the statistical analysis in Chapter 4. I am thankful for all their contributions and help.

All other work conducted for the thesis (or) dissertation was completed by the student independently.

Funding Sources

This work was made possible in part by the National Science Foundation under grants IIS-074742, IIS-1247126, and IIS-1528044.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Researchers strive to make new contributions. Their work, which is generally referred to as "doing scholarly research", requires perusing vast information, including prior work in relevant fields, and developing coherent ideas, to derive research products. Discovering and digesting prior work is important in scholarly research, because prior work collectively represents a body of existing knowledge and intellectual achievements, and functions as a foundation for developing new contributions and communicating with the academic community. The importance of prior work is easily seen through the fact that, every scholarly publication includes a list of references to prior work, and a discussion of how they relate to the present work.

Previous information science research addressed how scholars seek [Bates, 1989; Belkin et al., 1982; Blandford and Attfield, 2010; Marchionini, 1997, 2006; White et al., 2006a; Wilson et al., 2010], read [Marshall, 1997, 1998; Marshall and Bly, 2005; Marshall et al., 1999; Schilit et al., 1998], sensemake [Pirolli and Card, 2005; Russell et al., 1993], manage [Boardman and Sasse, 2004; Jones, 2007; Jones et al., 2001; Whittaker, 2011], and archive [Marshall, 2008] information. Based on the findings, many tools, such as digital libraries, search engines, and bibliography managers, have been designed and developed to support these tasks.

However, researchers do more than these. Their work is also creative. The bottom line of scholarly research is to conceptualize and develop new, sometimes outrageous, but valuable ideas. Researchers need to discover gaps in the body of knowledge, articulate research problems, develop new methods, interpret study results, and formulate new theories, in order to make new contributions. We use *scholarly research ideation* to refer to the tasks in which people develop new ideas for scholarly research, and the creative processes involved in these tasks. Scholarly research ideation is a key component of what researchers do. Prior work plays a crucial role in scholarly research ideation: it is the foundation on which researchers build up new ideas, yet sometimes it can suppress ideation by causing fixation, a mental set counter-productive for creativity.

While information science research on how researchers work with prior work is valuable and

strong, we identify a gap in understanding and supporting researchers' practices. Creativity, often manifested as personal curiosity, learning, and growth [Beghetto and Kaufman, 2007; Kaufman and Beghetto, 2009], is not the primary focus in this line of research. The creative nature of what researchers do demands that investigations into their practices and experiences should ground in creative cognition, the empirical study of cognitive processes involved in creativity [Finke et al., 1992]. We found that the gap in addressing the creative nature of scholarly research ideation tasks sometimes takes form as experience breakdowns [Winograd and Flores, 1986] with tools researchers use. For example, breakdowns occur when researchers writing new articles find it difficult to think across and connect multiple prior works they have read before. Thinking across and connecting remote concepts is vital for developing new ideas [Wilkenfeld and Ward, 2001].

We argue that, to gain a deeper understanding of researchers' practices and experiences, and better support their needs for engaging in creative endeavors, we have to take an interdisciplinary perspective, integrating valuable ideas and findings from multiple fields, including creative cognition and information science. Creative ideation should become the focus of investigation, an umbrella encompassing and connecting other activities researchers engage in, since ideation is at the center of what researchers do.

This dissertation focuses on understanding and supporting researchers' creative ideation processes in scholarly research, specifically their engagement with prior work in ideation processes. We take an interdisciplinary, ideation-centered perspective in framing a new study of the scholarly research ideation practices and experiences of established principal investigators. Research questions motivating our investigation include: How do researchers work with prior work while developing ideas? What works in their practices, and what does not? How can we support them in their creative tasks? Our interdisciplinary perspective enables us to connect activities that were previously studied separately, such as seeking and management, and how these activities contribute to ideation. In particular, we identify *curation*, the creative conceptualization and organization of new contexts to recontextualize information and create new meanings, as an important component in achieving scholarly research ideation. We then applied these findings in the derivation of a new

model of scholarly research ideation, incorporating concepts from multiple fields.

We explore new opportunities to support scholarly research ideation. To address researchers' needs for thinking across and connecting multiple prior works, we designed and developed a new dynamic interface which presents interconnected web semantics in an integrated context. By *web semantics*, we mean computational representations—of information that is found on the web—which capture meaning involving people's experiences of 'things' of interest. Web semantics can provide significant attributes, descriptions, representations, relationships, contexts, etc, allowing people to think about, make sense of, and act on the underlying things of interest. For example, bibliographic information a researcher finds in a digital library is a form of web semantics: it provides meaningful context to the researcher, so that she can evaluate its relevance, make decision on if and how to get it, collect for future reference, and share to collaborators or students as a *boundary object* [Star and Griesemer, 1989]. By presenting bibliographic information as concise, consistent, and usable web semantics, the interface is found to support scholarly research ideation while the researcher seeks prior work in our study.

The new interface is just a start. The technology used to build the new interface supports modeling, extracting, and presenting not only bibliographic information, but also general *web semantics*, which are structured data found on the web, providing information about things of interest to the user. A type system is used to integrate reusable data models, extraction rules, and presentation hints, in order to support the dynamic nature of the interface. We envision more such dynamic interfaces supporting scholarly research ideation to emerge.

Contributions of this research include:

1. Assembling perspectives across disciplines,

2. Deriving findings from qualitative analysis of researchers' practices,

3. Integrating sensitizing concepts and findings into a new model of scholarly research ideation,

4. Developing a dynamic exploratory browsing interface to support scholarly research ideation,

5. Using a type system to integrate modeling, extracting, and presenting dynamic web semantics.

This research draws on and connects sensitizing concepts from multiple fields, which we review in Chapter 2. Chapter 3 presents the qualitative study with established principal investigators, and derives a new model for scholarly research ideation. In Chapter 4, we demonstrate a new interface for citation chaining, show that it supports scholarly research ideation, and use the new model to explain the result. Chapter 5 describes the underlying technology for building the new interface, and how it can be used to build similar interfaces to support ideation tasks. We then discuss findings and implications in Chapter 6, before concluding in Chapter 7.

2. SENSITIZING CONCEPTS *

This research draws on and connects sensitizing concepts from multiple fields, including cognitive science, information science, art, web science, and programming languages. This chapter discuss how these concepts provide a basis for our investigation, and how they relate to our work.

2.1 Creative Cognition [Geneplore]

Creative cognition conceives of creativity as a product of many mental processes. The same processes can lead to creative or noncreative outcomes. The creative cognition approach follows the the *family resemblance* principle: while most creative endeavors result from most mental processes concerned by the approach, no single process is necessary or sufficient. The family resemblance principle's openness enables the creative cognition approach to address diverse forms and aspects of creativity [Finke et al., 1992].

Geneplore is a model of creative cognition [Finke et al., 1992]. The model connects generative and exploratory cognitive processes.

In generative processes, one constructs *preinventive structures*, i.e., internal (mental) representations possessing preinventive properties. Examples of generative processes include association, in which people relate elements, and synthesis, in which people combine elements to form a new whole. *Preinventive properties* are characteristics found to promote creative discovery. Examples of preinventive properties include ambiguity, meaningfulness, and emergence. Ambiguity is the quality of enabling being understood in two or more ways [OED, 2017], which, in turn, avoids over-specifying. As a preinventive property, ambiguity supports novel interpretations [Finke et al., 1992]; this explains the mechanism underlying more recent HCI findings, about how ambiguity serves as a resource for design [Gaver et al., 2003]. Meaningfulness indicates a significant quality,

*Part of this chapter is reprinted with permission from the Related Work section of "Metadata Type System: Integrate Presentation, Data Models and Extraction to Enable Exploratory Browsing Interfaces" by Yin Qu, Andruid Kerne, Nic Lupfer, Rhema Linder, and Ajit Jain, 2014. In *Proceedings of the SIGCHI Symposium on Engineering Interactive Computing Systems*, pages 107-116, DOI: `https://doi.org/10.1145/2607023.2607030`.

purpose, or value. Emergence refers to phenomena in which unexpected features and relationships become apparent only after a whole is formed [Finke et al., 1992].

In exploratory cognitive processes, one looks for opportunities for developing preinventive structures into creative solutions. Examples of exploratory mental processes include interpretation, reflection, and evaluation. In interpretation, people contextualize and derive new meanings. Through reflection, people look back on what they know and what they do, to articulate previous implicit understandings [Schön, 1992]. By evaluation, we mean the act of examining how a particular preinventive structure is valuable, in a situated activity context.

Exploratory cognitive processes may directly lead to a creative idea applicable to one's task at hand. Otherwise, one modifies the developing concept represented by preinventive structures, by focusing it or expanding it, in preparation for another iteration of generative processes. When one gets stuck in a counter-productive mental set that impedes generation of new ideas, *fixation* arises [Smith and Blankenship, 1991].

Geneplore processes are contextually situated. Context is not a fixed set of surrounding conditions, but is wide and dynamic; human cognition is a part [Hutchins, 1995]. For example, personal experiences and knowledge are often valuable in making meaningful associations and interpretations, which lead to creative ideation. Information in the environment, such as task constraints, can affect ideation in both positive and negative directions [Finke et al., 1992]. We draw on Geneplore to understand participants' creative ideation in scholarly research, and incorporate it as a principal component of the new interdisciplinary model.

2.2 The Four-C Model of Creativity

Researchers position creative acts on a spectrum of impact levels: transformative interpretation ("mini-c"), everyday innovation ("little-c"), professional expertise ("Pro-C"), and societal accomplishments ("Big-C") [Beghetto and Kaufman, 2007; Kaufman and Beghetto, 2009]. At one end of the spectrum, *mini-c creativity* refers to personally novel, meaningful, and transformative interpretation and learning, of experiences, actions, and events. Little-c creativity is said to impact peers, friends, and family; this is naturally extended to social networks. Pro-C means impact on a

field, while Big-C creativity implies changing the world.

While the goal of scholarly research is to achieve Pro-C and Big-C, as is typical of creative processes, high-impact research often springs from mini-c steps, e.g., gaining understandings of relevant topics, reflecting on the status quo, discovering new connections, and shifting perspectives. The present research addresses ideas developed on the mini-c impact level, and the processes through which experienced researchers grow these initial mini-c ideas toward interpersonal, professional, and even societal significance. The mini-c perspective of creativity enables investigating processes of scholarly research ideation, across levels of impact, as connected.

2.3 Working Memory Limit

People engaging in complex cognitive tasks, including scholarly research ideation, face the fundamental limitations of human cognition. Studies found that working memory, which is used to store and manipulate information to support complex cognitive tasks, is limited to about 4 chunks [Cowan, 2001; Miller, 1956]. This becomes an obstacle for human creative endeavors, including scholarly research, which often requires thinking of and connecting multiple elements.

2.4 External Representations

External representations, previously defined as "knowledge and structure in the environment, as physical symbols, objects, or dimensions, and as external rules, constraints, or relations embedded in physical configurations" [Zhang, 1997], are found to enhance cognition in performing many kinds of tasks. Zhang and Norman analyzed how representations of numbers support solving mathematical problems [Zhang and Norman, 1995]. Hutchins showed how people interact with external representations, such as charts, and engage in external processes, such as manipulating artifacts and communicating with other people, to accomplish the complex and situated task of navigating at sea [Hutchins, 1995]. Neuwirth and Kaufer found that external representations support synthesis in writing [Neuwirth and Kaufer, 1989]. Suwa and Tversky found that external representations, which can be interpreted in multiple ways, support generation and development of design ideas [Suwa and Tversky, 2002]. Norman used notions of "knowledge in the head" and

"knowledge in the world" to respectively refer to internal and external representations. He suggested that good design should use external representations to help users recognize, rather than remember, how an artifact can be used [Norman, 2002]. Kirsh showed multiple ways external representations support people in thinking [Kirsh, 2010].

In his influential work, Hutchins said: "... heavy interaction of internal and external structure suggests that the boundary between inside and outside, or between individual and context, should be softened... Instead of conceiving the relation between person and environment in terms of moving coded information across a boundary, let us look for processes... among elements of a system that includes a person and the person's surroundings." [Hutchins, 1995, p. 288]. Scaife and Rogers reviewed studies and models, and concluded that the interplays of internal structures and external representations are the key for understanding how graphical representations enhance cognition. They further suggested that these interplays are complex and cyclic [Scaife and Rogers, 1996].

Leading research on the embodied nature of human cognition shows that our thinking is deeply rooted in how we perceive and interact with the environment, using our body [Glenberg, 2015; Glenberg and Kaschak, 2002; Lakoff and Johnson, 1999; Merleau-Ponty, 1962; Tversky, 1993; Varela et al., 1992]. Sensorimotor experiences are found essential for cognitive processes, including those seemingly abstract, such as language comprehension and goal understanding [Glenberg, 2015]. Linguistic constructs involving spatial relationships play a key role in how we make all sorts of comparisons and form analogies [Lakoff and Johnson, 1999]. External representations thus are constituents of embodied cognition, as they are perceived and interacted with by people. *Distributed cognition* builds on embodied, by continuously incorporating social relationships and the world, manifested through external representations, in a broad, connected systems view [Bateson, 1972; Wiener, 1961] of what people operate on and act through, and how [Hollan et al., 2000].

2.5 Curation

The concept of curation originated in art practice, and grows as art practice evolves. In 1917, Marcel Duchamp developed the notion of *found object* by taking a urinal and exhibiting it, in a new context, as an artwork, with a new title, *Fountain*. Ironically, Fountain, which was exhibited under protest by organizers, became the exhibition's most famous piece. This act of choosing and exhibiting—i.e., of *recontextualization*—transformed the object's meaning and significance [Lippard, 1971]. Subsequent Dada [Lippard, 1971] and Surrealist [Breton, 1924] artworks advance the notion of found object by appropriating everyday artifacts as material for assemblage. Assemblage is an art-making act, in which found object materials are fastened together, showcasing the duality and tension between the original and resulting contexts [Seitz, 1961].

In other fields, the concept of curation is widely used to refer to collection, preservation, storage, and management of data. In e-science, researchers "curate" scientific data, including results generated from experiments and studies, documentations, codes, and metadata [Abbott, 2008; Gray et al., 2002; Higgins, 2008]. The purpose is to make scientific data available, accessible, and reusable to the research community, in the present and future. In personal information management, "curation" refers to keeping, managing, and exploiting familiar personal information, focusing on how people organize, navigate, and search for information [Whittaker, 2011].

While art practice evolved, the concept of curation, itself, developed from passive organization of and caring for a collection of artworks for public display, to active specification and assemblage of artworks for exhibition. Curators became directly involved in production and presentation of artworks, and thus as creators of art. They appropriate, assemble, and recontextualize creative works to develop new ideas. Curation becomes a means for conceptualizing how fields and their contexts are understood [O'Neill, 2012]. In this sense, curated works function as found object *material* [Giaccardi and Karana, 2015; Miller, 2004]. The act of curating, in turn, contributes to its fields through presentation and discussion [O'Neill, 2012]. Curation emerges as a distinct mode of discourse and framework for interaction across disciplines.

A popular practice, today, is to derive new meanings and tell stories by collecting, sharing,

and commenting, in social media and generally on the web, using found information [AlNoamany et al., 2017; Linder et al., 2014; Zhao and Lindley, 2014]. It makes particular sense to refer to this practice using the concept of curation, because, beyond finding, collecting, and managing information, "curation" explicitly connects to production of meanings through recontextualization. We draw on the active, transformative, and creative connotations of the term *curation*, viewing curation as a "contact zone" of information, ideas, and minds, in which new meanings can emerge [Dallas, 2016]. In particular, we find that researchers curate prior work, findings, thoughts, and other kinds of information, to help develop new ideas.

2.6 Information Seeking

Information seeking is a broad term, referring to activities that one purposefully engages in to change her state of knowledge [Marchionini, 1997]. Seeking information, including that of relevant prior work, is an important part of doing scholarly research. Developing understandings of relevant prior work is often essential for deriving new contributions.

Search and browsing are two information seeking strategies [Marchionini and Shneiderman, 1988]. Early work conceptualized search as a single-shot process, in which a computational system accepts a query—presumably representing the user's information needs—and computes a score for each pre-indexed document, measuring the document's relevance to the query [Bookstein and Cooper, 1976]. Belkin pointed out, early on, that a single query can rarely capture a user's situated, evolving information needs. Rather, he developed the notion of "anomalous state of knowledge" to describe situations in which the user recognizes a want for information, without necessarily knowing exactly what is needed [Belkin et al., 1982]. Bates used *berrypicking* to describe the incremental gathering of scattered information through iterative searching, browsing, discovering, learning, reformulating queries, and reevaluating potential leads [Bates, 1989]. *Exploratory search* refers to processes in which people investigate and learn through iterative search; their understandings and information needs evolve in the course [Marchionini, 2006; White et al., 2006a; White and Roth, 2009; Wilson et al., 2010].

Browsing is an informal and opportunistic strategy for information seeking [Marchionini, 1997].

Scanning through books on a shelf is a form of browsing. The World Wide Web and digital libraries streamlined browsing. Previous research used *exploratory browsing* to refer to processes in which people traverse hyperlinks to seek novel information, as they use hypertext to investigate topics, explicitly addressing the serendipitous and situated nature of browsing on the web [Jain et al., 2015; Qu et al., 2014].

Kerne used the etymology of "browse" to identify its spontaneous, whimsical nature; the term originally referred to how deer and cows spontaneously pursue the tender parts of rough plants [Kerne, 2001]. We build on this concept of "browse" to frame a notion of *evolving interests*, which characterize people's situated states of desiring particular information. What one finds *interesting* is situated and personal. An interest may connect to a specific information need [Chi et al., 2001], an anomalous state of knowledge [Belkin, 1980; Belkin et al., 1982], an unexpected discovery worth exploring [Kerne and Smith, 2004], or even a curious idea that one simply cannot help pursuing. Interests may be, in any combination, considered and impulsive.

Note that, both containing the word "exploratory", exploratory mental processes (in Geneplore) and exploratory search / browsing (in information seeking) involve investigating the unfamiliar. In Geneplore's exploratory mental processes, one investigates preinventive structures and properties *in the mind*, to develop ideas. In exploratory search / browsing, one instead investigates information *in the world*, to learn and discover. Despite this contrast, processes in the mind and in the world are integrated and interweaved in practice, working together to facilitate developing ideas.

2.7 Studies on Researchers' Practices

Ellis developed a behavioral model of how researchers seek prior work [Ellis, 1989; Ellis et al., 1993]. In particular, *citation chaining*, the act of recursively finding articles cited by or citing the current one, to discover more interesting work, is found a popular method. We consider citation chaining as a form of exploratory browsing.

Marshall studied how students make annotations while reading books [Marshall, 1997], and how researchers collect and archive materials used in research [Marshall, 2008]. Findings show that making annotations is an integral part of reading, helping readers understand the material, and

it is not uncommon for researchers to make annotations, in the form of comments, summaries, and notes, along with bibliographic resources they collect into their personal repositories. Making annotations and collecting bibliographies to form personal repositories are actions of curation. Our work builds upon these studies, and further investigates how these actions of scholarly curation support scholarly research ideation.

2.8 Web Semantics and User Interfaces

We use *web semantics* to refer to structured data found on the web providing details about things of interest to the users, as well as its computational representations affording users' actions (including but not limited to reading, thinking, collecting, clicking, and sharing). It is also called *metadata* in many contexts. In this work, we use the two terms interchangeably. For example, web semantics about a product being sold online would include product name, description, price, photos, and reviews.

Because metadata contains important information that is of interest to the users, interfaces often need to obtain and present metadata to support user tasks. For researchers, the most frequently used metadata is bibliographies, for example, of prior work they build upon, cite as supportive evidences, and compare to. Interfaces supporting researchers work with prior work must also support obtaining, presenting, managing, and using bibliographic metadata. The present work uses a type system, to integrate reusable data models, extraction rules, and presentation hints, to accomplish this. The runtime obtains metadata from regular websites through extraction, thus not relying on websites to publish formal semantics in special formats. Reusable presentation hints help generate metadata presentations in user-friendly forms. In this section, we review the space of work related to obtaining and working with metadata, and contrast prior approaches with ours.

2.8.1 Metadata on the Web

The Semantic Web [Antoniou and van Harmelen, 2004] effort develops standards and techniques to represent, query, and process metadata. RDF [W3C, 2004a] is the primary information model. It represents metadata as *triples*, each consisting of a subject, a predicate, and an object.

RDF can describe complex, interlinked metadata and relationships. However, many useful web sites and services, e.g. Google Search, Amazon, ACM Digital Library, and Twitter, do not publish RDF. RDF-S [W3C, 2004b] and OWL [W3C, 2009] are Semantic Web technologies that specify metadata schemas using RDF. The focus is on inference rather than presentation. In consideration of contemporary web programming practice, we observe that presently out of the 21,100 APIs indexed by `http://programmableweb.com`, a search of "RDF" shows only 63 API results. Thus, the Semantic Web representation of types and data seems unpopular with web developers. This problem is not new [Ankolekar et al., 2007].

Microdata [W3C, 2012] embeds metadata into HTML pages using attributes denoting types and properties, making it easier for websites to publish formal semantics. Major search engines have been collaborating on a set of standard semantic types described in microdata, at `http://schema.org`. However, like RDF, many useful web sites, including Amazon and the ACM Digital Library, do not publish microdata.

2.8.2 Extracting Metadata

To overcome the scarcity of metadata on the web, prior systems extract metadata, for the user to collect and view.

Web Summaries [Dontcheva et al., 2006] is a browser extension that allows users to create extraction patterns, extract metadata from web pages, and see collected metadata in different views. In a 10-week study [Dontcheva et al., 2008], extracted metadata was found useful for both transient and long-term user tasks. Users liked a functionality called that takes a hyperlink on a page, extracts metadata from the destination page, and brings it into the context of the current page. They valued directly accessing linked metadata from within an initial context.

Piggy Bank [Huynh et al., 2005] extracts metadata from web pages using a browser extension, stores it in an RDF database, and provides a faceted browsing interface. Exhibit [Huynh et al., 2007] allows the user to publish metadata in a special JSON-based format in a faceted interface. Presentation is templated and customizable.

Marmite [Wong and Hong, 2007] and Vegemite [Lin et al., 2009] let end users create browser

based metadata extraction scripts, or *mashups*. However, studies showed that users without programming skills experienced difficulty in authoring mashups. Thus, the applicability of these tools to general, unskilled users is unclear.

Clui [Pham et al., 2012] provides a browser plugin for users to collect metadata from the browser, represented with types called "webits". The type system lacks inheritance or polymorphism, and so is limited.

The present metadata type system actively extracts metadata from regular web pages. Different from scrapers that extract individual metadata records from open browser windows, the type system supports extensibility by enabling developers to reuse data models and presentation hints, through an object-oriented programming language, as well as to reuse types, such as scholarly article, across different information sources. Further, this research addresses presenting *linked* metadata in one context. while making relationships visible, to support exploratory browsing.

2.8.3 Exploring and Presenting Metadata

mSpace is a faceted browser for exploring a fixed repository of knowledge in the form of a metadata collection [schraefel et al., 2003]. The user can re-order facets (dimensions) to reorganize presentation. When the user hovers over a facet label, mSpace shows associated snippets extracted from documentsl, bringing limited information into context. mSpace requires the knowledge to be encoded in RDF in advance [Smith, 2011], preventing exploratory browsing of newly encountered information.

CS AKTive Space presents a UK Computer Science research metadata collection [Glaser et al., 2004]. The interface displays search results in a faceted list, supporting column sorting and preview cues, like mSpace. When the user selects a research group, person, or publication, details are shown beneath the faceted list, in the same page, maintaining context. However, only one detailed item can be shown at a time. Metadata is collected through *ad hoc* programs translating data to RDF, called *mediators*. They have been used "predominantly for large, comparatively static data sources" and "high-value data sources of general interest to the community" to populate the system with enough data, implying a scarcity of RDF data in the domain. Since knowledge acquisition

precedes interaction, the ability for serendipitous browsing and exploration is limited. The system only addresses information in one domain, not an open-ended set of heterogeneous sources.

PGV [Deligiannidis et al., 2007] visualizes interconnected metadata in RDF as a graph; nodes are entities and edges are relationships. The user can expands linked nodes incrementally. The Atom Interface [Samp et al., 2008] improves visual presentation of such a graph using circles. X3S [Stegemann et al., 2012] reconstructs RDF query results in XML, which is further transformed to HTML styled with CSS for presentation. These interfaces operate on prepared RDF datasets, and thus do not support open exploratory browsing.

Tabulator [Berners-Lee et al., 2006] is a generic browser for linked RDF data. Its outliner mode shows metadata in a manner similar to MICE. Tabulator supports serendipitous browsing, differentiating from prior RDF interfaces. It is more generic. When the user expands a field, and the field is a link to another metadata record, it actively dereferences the link and shows connected metadata in the same context. The authors emphasized such serendipity, since it supports "re-use of information in ways that are unforeseen by the publisher, and often unexpected by the consumer". However, Tabulator's scope is limited by the scarcity of RDF data on the web. The absence of type-based presentation hints leaves issues of metadata's cognitive load un-addressed.

Parallax [Huynh and Karger, 2009] enables "set-based browsing". The user starts with a set of metadata records, connected by facets. However, when browsing across facets, direct presentation of context is lost. The user views metadata linked to the current set in a new viewport. To ameliorate, Parallax maintains a linear trail of previously browsed sets. However, browsing and exploration processes may not be linear [Greenberg and Cockburn, 2002; Klemmer et al., 2002]. Parallax works with a prepared dataset, available at `freebase.com`. Users thus cannot explore live information outside the prepared dataset, such as ACM Digital Library papers.

2.9 Type Systems

In programming languages, types are used for organizing computational entities that share certain common properties [Mitchell et al., 2003]. Modern programming language compilers and interpreters can conduct *type checking*, to ensure that each and every entity is used correctly ac-

cording to its type, and report errors if not so. Type information and type checking make programs easy to read and understand, and less error-prone.

Many programming languages support creating new types by combining existing types, as one of the methods for building abstractions in constructing complex systems. After a type is created and named, it can be referred to by name, e.g., in function signatures, rather than repeating its internal structures again. As types become more and more complex, in many cases, reusing a type becomes necessary. Inheritance is a way for reusing a type. When Type A inherits from Type B, it means Type A automatically gets the internal structures of Type B, potentially with additional, new internal structures.

In programming languages, polymorphism refers to constructs that can take on different types as needed. In particular, subtyping polymorphism enables a variable of a base type to take on values of subtypes. In object-oriented programming languages, thanks to dynamic lookup, subtyping polymorphism allows the same general method call to behave differently on objects of different subtypes, so that functionalities can be extended without changing the call sites. This allows for great extensibility and flexibility.

In the Semantic Web, types are also used to categorize semantic information. With RDFS or OWL, one can specify complex metadata types, reuse data models through subtyping, and organize types into ontologies.

The present research develops a metadata type system to support dynamic extraction and presentation of web semantics, to address user needs in scholarly research ideation activities. We will compare our approach with previous uses of types and type systems in the discussion.

3. A MODEL OF SCHOLARLY RESEARCH IDEATION

To understand how researchers engage with prior work in performing scholarly research ideation tasks, we conducted a qualitative study with 14 established principal investigators. While researchers also generate insights and new ideas from many sources, e.g. when engaged with raw materials and environments [Caro, 2019], this study instead focuses on the moments when participants look for or use prior work to support *mini-c* creative ideation [Beghetto and Kaufman, 2007; Kaufman and Beghetto, 2009], i.e. personally novel interpretations, thoughts, reflections, and inspirations about a research topic or product.

We derive findings through a grounded theory methodology, and develop a new, interdisciplinary model of scholarly research ideation, connecting findings and sensitizing concepts. The goal of the new model is to provide a new, formal understanding on participants' scholarly research ideation practices and experiences, that focuses on creativity. The scope of the present study is limited by its fairly small n, and the breadth of the study participants, who are concentrated in fields of human-computer interaction, computer science, and interactive media. We expect that they are relatively computer literate. Most are science-oriented, though a few lean toward art, design, and humanities. Further, our study is focused on how researchers use prior work, rather than raw data, in processes of ideation.

3.1 Methodology

Charmaz articulates an iterative empirical qualitative methodology for constructing grounded theory, involving: drawing on sensitizing concepts, gathering rich data, initial coding, focused coding, categorization, and conceptual refinement [Charmaz, 2006]. We apply this methodology to investigate scholars' practices of building on prior work in research ideation. We conducted semi-structured interviews to gather qualitative data, and analyzed data through initial and focused coding.

For the interviews, we recruited 14 principal investigators from 7 renowned institutions across

ID	Gender	Citations	Discipline - Research Areas	Seniority as a PI
R1	F	< 100	Visualization - Art	Junior
R2	F	4000	Computer Science - Systems	Senior
R3	F	200	Visualization - HCI	Junior
R4	M	9000	Computer Science - Bioinformatics	Senior
R5	M	2000	Computer Science - HCI, Design	Senior
R6	M	10000	Arts & Media - HCI, Information Visualization	Senior
R7	M	12000	Psychology - Creative Cognition, Memory	Senior
R8	M	3000	Computer Science - HCI, Social Media	Senior
R9	F	9000	Computer Science - HCI, Information Retrieval	Senior
R10	M	n/a	Computer Science - HCI, Virtual Reality	Senior
R11	F	n/a	Computer Science - HCI, Creativity	Senior
R12	F	1000	Information Systems - HCI, Creativity, Art	Senior
R13	F	< 100	English - Digital Humanities	Junior
R14	F	5000	Computer Science - Systems	Senior

Table 1: Study participants. Citation count is from the participant's Google Scholar author page, rounded to nearest hundred or thousand. "n/a" indicates that the participant does not have a Google Scholar author page, thus we were unable to obtain a citation count. Citation count data is up to July 2017. We regard researchers, who have held a position of associate professor or equivalent for more than 3 years, as senior.

the world (Table 1). Most participants are senior researchers (e.g., full professors), while some of them are relatively junior (e.g., assistant professors). Most participants work in fields including HCI, Digital Humanities, and Cognitive Psychology. Our participation in these fields facilitated developing rapport with participants and contextual understandings of their practices. We conducted a single-session, semi-structured interview with each participant. Each interview took 30 to 60 minutes. Seven interviews were conducted via online video or audio chat; others offline, in person.

In the interviews, we asked participants about their practices of developing research ideas, including finding and building upon prior work. Sample interview questions included: (1) How do you conceptualize and develop research ideas? (2) How do you find, choose, and use relevant prior work in research? (3) What problems have you encountered with finding and using relevant prior work? (4) How do you reflect on your practices?

For in-person interviews, we also asked participants to perform daily tasks, explain procedures, and show physical / digital artifacts in their working environments, based on what they mentioned in their responses. With their permission, we collected some photos or screenshots of relevant artifacts. Following an iterative study design strategy, we refined our guiding questions and interview procedure according to findings discovered in the process. For example, we followed up with several senior researchers about their processes of developing ideas with regarding to one of their most influential works, to ground the study in specific contexts.

Interviews were transcribed to facilitate analysis. Three researchers were engaged in initial and focused coding. We began by taking notes of noticeable statements or artifacts, with which we formed initial codes. We then identified interesting, recurring themes. To derive clear and distinct focused codes, three researchers regularly met in person, to examine and compare each other's initial codes. Each code was validated across gathered data, and reflected upon, until consensus was reached on its validity and significance. Similar codes were combined together. Relationships between codes were explored. Focused codes emerged as we gained understandings of the data. We repeated this process to refine codes, until no new codes emerged. We organized focused codes into categories, to synthesize and derive findings.

3.2 Findings

Based on perspectives framed by sensitizing concepts, we conducted qualitative data analysis and derived findings on participants' practices in performing scholarly research ideation tasks. In most cases, the task involves conceptualization of ideas, conduction of research, derivation of findings, and synthesis into research products (e.g., scholarly articles). We first present findings on how participants generate initial ideas. Then, we look into their practices of developing initial ideas and conceptualizing well founded and novel research projects, in which prior work plays an essential role. The findings further show that participants' scholarly research ideation practices are deeply situated in multiple contexts. We then present findings on how participants use methods of curation to address the situatedness of what they do.

3.2.1 Generating Initial Ideas

Participants' initial ideas for research projects come from many sources. For example, R2 starts with observing scenarios in which computer systems fail or underperform, which she calls "technical pain points". R5 and R7 emphasize the importance of their own personal interests. R6 starts by asking how one can use visualization to make a task easier for people, or enable a wider population, to perform it. Many participants were inspired by discussions with colleagues.

3.2.2 Ideation with Prior Work

Growing initial ideas into well founded, novel research projects requires finding and building upon prior research. All participants found prior work essential for scholarly research ideation. First of all, to make new contributions, and to persuade the research community to accept claims of new contributions, researchers need to evaluate and position their work in a context that becomes constructed by prior work.

> Q1 * R5: [Conceptualizing a research project] is a matter of being aware of what is going on in the community, having the interests, and being able to position what we are doing with respect to what other people are doing.

To preliminarily evaluate the novelty and impact of a potential research project, many participants conduct searches on prior work in the early stage of conceptualization.

> Q2 R3: First step [of investigating an unfamiliar topic] is to do a general Google Scholar search, to get an idea if [the topic] is worth pursuing as a project or research area, [and] to see briefly what people have done earlier.

In some cases, through seeking, reading, and making sense of prior work, one may find a lack of prior work that directly tackles a problem of concern, or, according to R2, a "hole" in prior work. The hole, or gap, becomes an opportunity for significant new work. For example, R9 has a paper which was the first to investigate the conjunction of an already well-known task, in a new

*We index each quote for subsequent referencing.

emerging setting, despite that neither the task nor the setting, by itself, was particularly new to the research community. The paper has been widely cited to motivate subsequent work on the same kind of problem. R7 found the then-time prevailing theories on a phenomenon uncompelling.

> Q3 R7: [About the phenomenon] I didn't like [the theories popular at that time] ... [because of] lack of evidence.

In some other cases, as R2 pointed out, a lack of prior work may reveal how one's initial idea is broken. This, in turn, motivates further iterations to develop the idea.

Concepts, methods, and results from prior work can be essential to one's ongoing research in multiple ways. One common case is to use prior work as components and building blocks. One work of R2's reviews prior concepts that were invoked in the formation of an emerging field. Prior concepts became components of this work. The contribution was then derived through explaining the origins of these concepts, discussing how they related to each other (e.g., the paper contains an ontology of related concepts), and providing insights on future research directions for what was then a young, emerging field. A seminal work of R7's uses experimental methods from prior work as building blocks, to perform studies for a different problem.

> Q4 R3: Sometimes when you have [an initial] idea that's vague and abstract, you can't concretize it... Then you read something and say, "Hey, that's the exact thing I wanted!"

For participants to associate concepts—from different fields or problem domains—is a common practice for developing new ideas. R4 brings theories and methods from one field to another. This results in not only successful publications, but also leads to more rigorous methods in the interdisciplinary field.

> Q5 R6 (when talking about one of his seminal work): All of these ideas existed outside of [the field of the work], but we incorporated and expanded on them... [The work] was a combination of taking ideas from [a list of different fields] and putting

them into the system for [a particular task]... Sometimes you combine ideas from what exists before. Sometimes you set up in new directions.

Another popular method for developing novel ideas is to making analogies across problems / domains / fields as sources of inspirations.

Q6 R6: I found a problem domain and [developed] a [new] solution. And from there, that solution is applicable to a number of different domains... In some ways, [a problem in a different field] is the inspiration for all these works, though it's much more than just [that original problem].

Q7 R7: There was another mystery [in a related research field]. And I thought, [the new mystery and the problem I have been working on] are kind of similar; they ought to be linked. The first thing to do is to go back to the literature. No one had ever [tested the link] before.

Q8 R7: When you try to solve a problem in one area or domain, which you don't know, you [can] reach for another problem in another domain [that] you already know about, that is similar, in the dynamics and relations, and say, "What if I can apply that other thing to this problem?". To be able to do that, you have to think abstractly enough to see how this problem is like that other problem.

Participants also warned about the risk of seeing prior work "too early", which could prevent one from trying new approaches—a state of fixation.

Q9 R6: I found that if you look at the prior work too early, it can stop you from making any progress, because it closes off your mind to other possibilities... You would have actually done it in a different way if you hadn't seen it. To me, the best approach is: [in early stage] you do a cursory examination of whether there is prior work that [is exactly the same idea].

Q10 R2: [A colleague professor] is against doing any search [in early stage]. [His method is] you do your work; once you're done, you look what other people did. He thinks that [seeing prior work too early] kills [creative] thinking... In his view, once you go and see prior work on the same problem, you put yourself in an incremental mode, and very rarely you go and think drastically differently.

Q11 R9: Even when you find work that you think is the same, if you go read it, you can often see that there is a useful, different perspective.

Finally, all participants find it important to appropriately cite relevant prior work when they write.

Q12 R2: You want to cite the papers that defined or first introduced the problem. Then the ones that you compare to. Then [papers] from groups looking at the same problem, but you can't compare to for some reason.

3.2.3 Seeking Prior Work

Because prior work is essential in participants' ideation processes, seeking prior work that contributes to the evolving generation of a new conceptual context—building blocks, components, and inspirations—becomes important. We investigated how participants seek prior work. Our findings are consistent with prior information seeking research [Bates, 1989; Belkin et al., 1982; Blandford and Attfield, 2010; Marchionini, 1997, 2006; White et al., 2006a; Wilson et al., 2010], but contextualized by a focus on how participants' seeking processes affect and contribute to their ideation tasks.

3.2.3.1 Chaining and Extracting: Seeking (mini-c) Novel Information

Participants use diverse techniques to seek prior work of interest, which is novel to them. For example, all participants are familiar with techniques of "citation chaining", i.e., using citation lists and indexes to find articles citing or cited by the current one. We observed an interesting form of

chaining: looking up an author's website for her bibliography or CV, to find articles by that author, which are previously unknown to the seeker.

Q13 R7: [Sometimes] I look at someone's website, to see their publications.

One technique participants use to stay up-to-date in their fields is to physically attend and participate in conferences. A digital alternative is to periodically browse top journals and conference proceedings, a practice of "extracting" [Ellis, 1989].

Q14 R12: I very rarely [have time to] go through journals or conference proceedings. Going to conferences is a good way to get a sense of what people are doing.

Q15 R1: I like to see all the proceedings. I go to the proceedings section and go through all papers. That helps to see the trends.

Some participants found social media useful for staying up-to-date with fellow researchers' work.

Q16 R6: I read a bunch of different blogs, follow a bunch of Twitter people who do interesting stuff.

Chaining and extracting are methods of exploratory browsing [Jain et al., 2015; Qu et al., 2014]. Exploratory browsing is inherently serendipitous and opportunistic [Blandford and Attfield, 2010; Marchionini, 2006]. One goal of exploratory browsing is to find information that is novel, unanticipated, even unheard of to the seeker / researcher before hand. Such information can be transformative to the seeker / researcher. Thus, we observe the significance of finding mini-c novel and valuable information—that is, information which is new to a particular participant—through chaining, extracting, and other exploratory browsing techniques.

3.2.3.2 Reading and Interpretation in Seeking: Iteratively Developing Ideas

People's goals for seeking include not only finding and understanding information per se, but also having ideas on its relationships to and implications for what people are doing. In our study, participants develop ideas on relationships and implications by constantly interpreting found prior work, in situated contexts. These interpretations, in turn, inform which potential leads to follow and which papers to read in-depth.

Q17 R7: [First] I'll read the title. If the title grabs me, I'll read the abstract. If the abstract grabs me, I'll read the first paragraph. If it still grabs me, I'll scroll down and read the results. If it really grabs me, I'll download and read the whole article.

Q18 R6: [When going through search results] I start off with relevant titles... look in the abstract to see: if this is not what I want or this is interesting.

Q19 R12: Usually, I read [the] abstract right away and then [pick] 2-3 papers [from the search results] that I want to completely read... depending on how crucial the piece of research is to what I am doing. [Whether] it is just a small part or what i am doing completely builds on it... [determines] how thoroughly I read it.

Q20 R3: Sometimes [the] paper title doesn't tell you much, but if you see an author that has worked in the field, that will be a flag.

Reading and making sense of found information is part of seeking. Participants learn about terms and topics from found prior work, and use them to extend their horizons.

Q21 R10: If I don't find [relevant prior work], I'll refine my search, to narrow it down... [I'll] add keywords if it's a topic. I look at search results, to see if I'm missing anything.

Our findings on participants' practices of interpretation are consistent with previous information science findings on anomalous states of knowledge [Belkin et al., 1982], berrypicking [Bates, 1989], exploratory search [Marchionini, 2006; White et al., 2006a; White and Roth, 2009; Wilson et al., 2010], and exploratory browsing [Jain et al., 2015; Qu et al., 2014]. As they gain understandings on found information and how it relates to their work, their *interests* evolve.

Participants' acts of interpretation and learning constitute mini-c creativity. Through generative processes, such as recontextualization, and exploratory mental processes, such as interpretation, mini-c ideas can grow into more impactful research projects and contributions.

> Q22 R3: You read about the theory, then ask, "How could you apply it to a particular problem of your domain?" Maybe a project can come up.

> Q23 R7: Sometimes the original concept is better than the later interpretations of it... Sometimes the original idea is flawed, but [a later] interpretation of it is better.

Prior work also plays an essential role in interpreting data generated from studies / experiments. Generated data needs to be interpreted to become contribution. Interpretation requires contexts. New ideas can only emerge when generated data is contextualized, e.g., compared to prior results, and interpreted in prior theories (to support or dispute them). In R4's case, through introducing new methods for analyzing experimental data, his team found a pattern. It only becomes interesting when they could not find any prior work reporting this pattern. After further investigation into prior work, they published a plausible interpretation of the pattern.

3.2.4 Situatedness of Ideation Processes

Participants do not mechanically follow prescribed, definite algorithms for developing new ideas. Rather, participants' ideation processes are deeply *situated* in their contexts of practice. Context is not a static setting, but a dynamic product of human activities [Dourish, 2004].

The findings we have presented so far clearly show that participants' ideation processes are situated in contexts of prior work. Making novel contributions requires a thorough knowledge of

prior work, e.g., what is known, what is unknown, and what has been done. Even when contributions are driven by results generated from experiments or studies, one needs to compare new results with prior ones, and make new interpretations, to justify how work is novel and useful. In many cases, researchers draw on remote fields and problems—which do not seem directly relevant, but exhibit certain structural similarities—to develop new ideas in their own fields. Thus, the dynamic context of (directly and indirectly related) prior works, along with researchers' understandings and thoughts about them, constitutes an essential context for scholarly research ideation.

Participants' ideation processes are also situated in their personal backgrounds and interests.

> Q24 R6: People find something that becomes their signature... I've got my basic principle, my hammer, and then I say, "can this principle work well in this task?"... I call it the "I've got a hammer" approach.

Writing a research paper is an ideation process based on composing an argument. This often involves synthesizing related prior work into a framing, which reveals gaps and shortcomings, thus enabling researchers to make new contributions. Synthesis means combining elements to form a complex whole [OED, 2017]. It thus is situated in, e.g., the writer's interests in research problems, understandings of related prior work, perspectives on involved fields, and study results. The writing process, itself, is often iterative; writers commonly experiment with multiple ways of combining words into sentences, and sentences into paragraphs. R5 summarizes his processes of synthesizing related prior work into a framing:

> Q25 R5: Part of the writing process... you have to have a perspective of how your research is novel. You form [an] argument... You look for data to support your argument, or prior work to base your argument in... When you [write about] related work, you are trying to give an overview to support the argument, and give reviewers enough background to understand what or how you are doing is novel. In that case, you come with broader categories for how to characterize the [research field].

Participants' information seeking processes, which often involve mini-c creative ideation, are also *socially situated*, e.g., in research communities that in which they participate.

Q26 R2: I don't look at anything about security or GPU. I don't pay very much attention to old papers... unless there is a very good reason.

Q27 R2: If you are submitting to a conference, you should look at the committee members. Because if they did something [relevant]... you have to know their angle.

Q28 R2: I understand what is published where... I know the groups. I know how they think, and what kind of things they do. I know their angles.

Q29 R10: I look at papers I cite, to see what they cite.

Q30 R11: Sometimes reviewers suggest [interesting prior work], and I'm grateful for that.

3.2.5 Curation: Supporting Situated Ideation

Because of its rich connections to production of meanings in art and people's use of digital information, we identify *curation* as a valuable term for referring to participants' practices of choosing, clipping, collecting, organizing, labeling, writing on, and critically thinking about prior work. The word "curation" can also refer to products created through these practices. We use curation as a sensitizing concept that effectively describes practice, even when it is not a word practitioners use.

We draw on prior theory to analyze how participants' curation products support their scholarly research ideation tasks, at two representational levels: *materials of elements*, which determine how individual found objects can be collected, interpreted, and interacted with, and *media of assemblage*, which prescribe how elements can be joined together to form a whole [Kerne et al., 2014; Webb et al., 2016]. We observed a particular, interesting form of curation integrated with seeking. We further identify breakdowns [Winograd and Flores, 1986] in participants' curation experiences with current tools.

3.2.5.1 *Materials of Elements*

Participants use diverse materials to represent elements as they curate found prior work. Some save PDFs to a local file system (e.g., R3, R7). Some collect bibliographic information as textual notes (e.g., R2, R6, R8, R14). Some create browser bookmarks and entries in their ACM Digital Library account binders (e.g., R10, R12).

Clipping is a particular situated material of elements in participants' curations. Marshall and Bly define clipping as the act of "intentionally saving portions of published material" [Marshall and Bly, 2005]. The word "clipping" can also refer to the resulting object. We extend *clipping* (verb)—emphasizing its found object function—as intentionally choosing and saving a portion of a physical or digital found information resource, to represent meanings significant to a scholar / curator, and recontextualize these meanings concerning the scholar's emerging research. In particular, we consider highlighting as a form of clipping. In highlighting, the reader intentionally

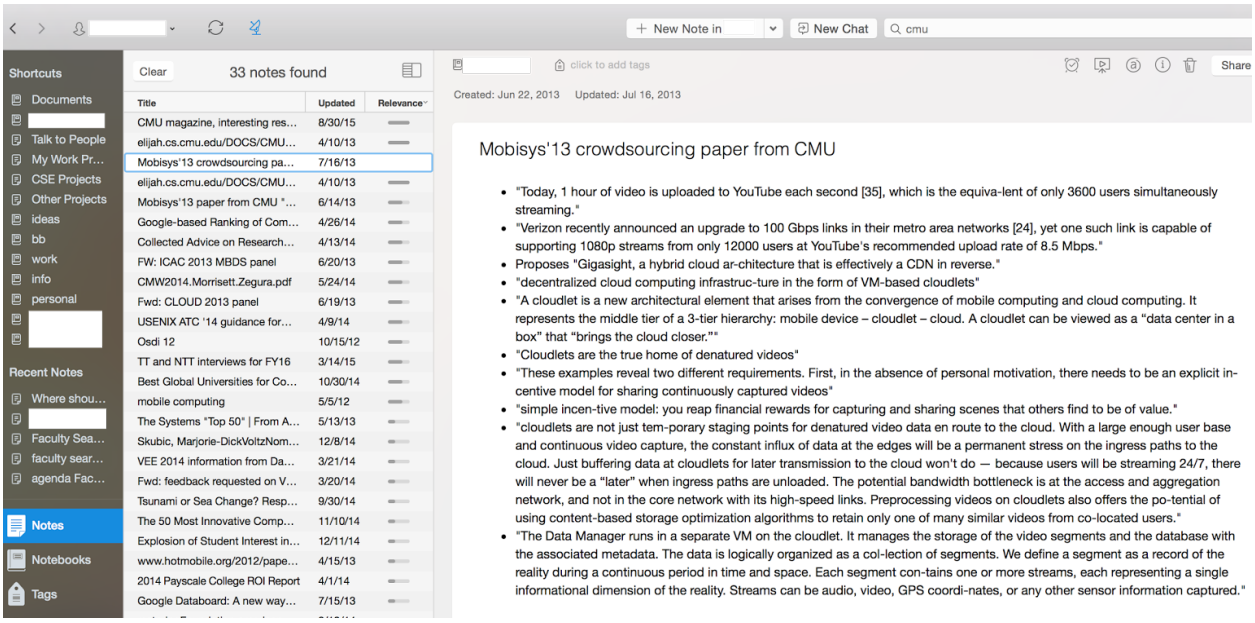

Figure 1: Screenshot of R2's prior work curation in Evernote, showing a list of prior work entries, and details of one. The list is filtered by term "CMU", and can be ordered by title, date added, or degree of relevance. For each prior work entry, R2 keeps chosen quotes based on her understandings.

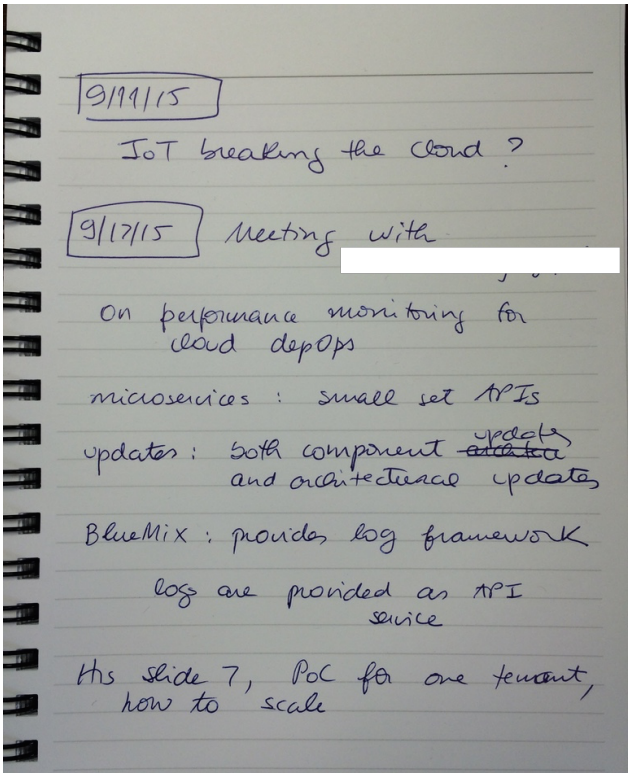

Figure 2: A photo of R2's handwritten notes in her Evernote library. Most notes in the photo represent thoughts on concepts related to her work.

chooses a chunk of text and saves a perceivable representation.

Many participants clip material from found prior work, to collect significant information and ideas. R7 makes screenshots of interesting portions from PDFs, and curates them in a Word document. R2 quotes sentences from a paper to summarize its key points, and curates with Evernote (Figure 1). R3 and R12 highlight interesting portions in a paper while reading. Others, such as R9, directly copy interesting portions of text from found prior work, and paste into a file or other textual medium to curate them.

Some elements are visual. R1 takes photos of interesting "visuals" encountered in daily life, draws pictures on paper, and keeps photos and drawings for inspiration. R2 diagrams on paper. She also takes digital photos of handwritten notes, which include thoughts on concepts related to her work, and curates them in Evernote (Figure 2).

3.2.5.2 Media of Assemblage

Media of assemblage, which can be physical and/or digital, determine how curated elements can be organized and exhibited, thus impacting how the curator and the audience think about the curation product. Common media of assemblage that participants use include linear lists, such as R2's notes (Figure 1), and a file system's hierarchy of folders. In dealing with folder hierarchies, duplicating files is common.

> Q31 R7: I download papers. I keep them in folders. I have folders in folders. When there are too many of them, I have [folders for] specialized topics. Some of them apply to more than one folders; I'll just have doubles.

> Q32 R3: A lot of papers are not limited to a single topic. I just duplicate the paper, for my sanity, so I don't have to look it up again.

Many participants assemble physical printouts of papers into piles. Each pile of papers represents a category.

> Q33 R1: This pile is "must read now". This pile is "maybe in a month or two". This pile is "not sure".

Some use more explicit visual media of assemblage. R1 uses paper to curate drawings, as inspirations. R3 uses a mind map app to explicitly show authors, papers, and relationships as a graph. She further states that visual representation of relationships helps her keep track of curated information.

> Q34 R3: I use mind maps... to have a visual representation of how everything is linked up, and get an overall idea of a domain. It helps me keep track of authors and papers. I can branch out [to add elements and make connections].

3.2.5.3 Curation in Seeking: Transient In-Browser Curation

We observed an interesting and popular practice. Participants keep many browser tabs open while seeking relevant prior work.

> Q35 R12: I open a tab for every resource that I find as a way to keep track of stuff. I keep all those tabs in one window. I might minimize that window while working on other things, and then zoom it back up when I'm back to working on that project. I only close those tabs when I'm done writing that paper and am sure I'm done with all those links and references. Sometimes it could be weeks.

> Q36 R6: At some point [in seeking prior work], I... end up with 40 to 50 tabs.

> Q37 R10: I tend to use the browser as my storage tool, rather than a specific reference [management] tool... Sometimes [browser tabs of found webpages] can be [open for] a few weeks.

> Q38 R12: [When] looking up a new topic... I might open a new tab for every resource I find, as a way to keep track of stuff... I'll keep [all these tabs] in one window. I might minimize that window when I'm working on other things, and [restore it] when I'm back to working on [the project]. I only close these tabs when I'm sure I'm done with all of the links and references... Sometimes it could be weeks.

We use *transient in-browser curation* to refer to this practice of informally choosing and collecting information pertinent to potentially useful prior work. Transient in-browser curation addresses researchers' need to track intermediate findings, which helps make sense of new information in context of other findings [Thomas and Cook, 2006], in their situated ideation processes.

3.2.5.4 Using Curation to Stimulate Ideation

Participants use what we identify as curation, as both process and product, to stimulate ideation. One method we observed is to write critical and reflective notes, while reading and curating found prior work. For example, R9 uses a postponed email to curate writings on relevant prior work and researchers she follows. She revisits these writings when working on new papers.

> Q39 R9: Answering High-Level Search Queries Using Human Computation, AAAI 2011: discuss how to use crowd computing to take a high-level user goal (such as "I want be healthier") and identify individual subtasks, with the aim of finding web search queries whose results help accomplish the high-level goal.

> Q40 R9: [A URL to a researcher's homepage] Intelligent task routing, reminiscence, pleasure in viewing old things. Maybe when tagging, people like to see their starred photos, motivates our use of stars as a way to select the photos to label.

> Q41 R9: To read: Shumeli, G (2010) To explain or predict? Statistical Science, 25(3). `http://arxiv.org/pdf/1101.0891.pdf` (Recommended by [another researcher's name])

Many of R2's writings (Figure 2) represent thoughts on concepts related to her work Like R9, she curates these writings. R3 keeps snippets and quotes for reuse. R13 has a folder for each ongoing paper, in which she curates PDFs of prior work, writings to incorporate in the ongoing paper ("notes"), and writings edited out of an ongoing paper for future use ("cuts"). After finishing a paper, some "cuts" become "notes". R10 revisits prior writings when working on a new paper.

> Q42 R3: I want to visualize all the notes while writing [a new] paper. I have a bag of snippets and quotes I want to use.

Q43 R13: [After completing a paper, in the folder] there will be a file of "notes", an article file, and "cuts"—stuff that didn't make it [to the completed paper] but I want to come back to. Sometimes it's my writing... sometimes it's from references.

Q44 R10: [When writing a new paper] I look at papers I'm [citing] to see key [citations] they used... and [prior work sections] from [my] previous publications.

Writing critical and reflective notes is an act of curation. For R9, writing helps her think critically and reflectively on concerned prior work, stimulating ideas on how prior work relates to, and can be used for her own work. For R13, R10, and R3, previous writings become roots and ingredients for ideas on how to write new papers.

Labeling is another form of writing, which involves acts such as naming a folder and tagging a pile (Q33). Labeling supports incremental shift of representations when people make sense of complex data and relationships [Kolko, 2007; Russell et al., 1993]. Through labeling, researchers categorize. Categorization is a basis for developing concepts [Lakoff and Johnson, 1999]. Thus, acts of labeling stimulate conceptualization of found information and relationships, and development of ideas.

Browsing curated prior work is another method for stimulating new ideas. This is consistent with creative cognition findings on provocative stimuli promoting generation of new ideas [Shah, 1998].

Q45 R3: Sometimes, I print out every single paper [relevant to the ongoing one] to have [them] laid out. It helps to have everything visually in front of you. When you're stuck, flipping through them helps trigger something.

3.2.5.5 Breakdowns with Current Tools

Winograd and Flores use the notion of *breakdown* to refer to moments when tools fail to support users' situated activities [Winograd and Flores, 1986]. With current tools, participants experience breakdowns in their scholarly research ideation activities. One breakdown happens in seeking, when participants get distracted and lost in the interconnected chains of prior work.

Q46 R3: Sometimes when reading a paper, I come across [an unfamiliar topic]. To understand the work properly, I start a new search for that topic. That search can become quite involved... I forget where I was [and] get lost.

Another breakdown happens in curation, when linear media for assembling elements limit participants in seeing nonlinear relationships.

Q47 R11: One problem in writing is seeing things visually. The whole process of writing is not a linear process. I use outline in Word, but that's a minimal tool. I'd like to be able to visually access the journals and my writings... see all the papers, and zoom in and out easily... [do] drawing and note-taking and everything.

When participants intensely engage in transient in-browser curation in their exploratory seeking processes, the limitation of the browser's linear medium of assemblage—long sequences of tabs, each representing an intermediate finding—corresponds to the breakdown of getting distracted and lost. It becomes impossible to even see all the titles of all the papers at one time, let alone clippings of relevant text. Further, because these curations are transient, it is easier to lose them.

Q48 R3: [After] I find a relevant paper, I look at the "cited by" list, and open [interesting citations] in new tabs. I search for [keywords learned from papers]. [After a while] it becomes very difficult to track.

Q49 R6: Initially most of my state is kept in a single browser session in the tabs. If I lose that browser session... I lose all those potential leads. And that has happened.

When it comes to using curation to support ideation, participants experience breakdowns when a prior work and writings about it are curated in separate places, thus cannot be looked at and thought of at the same time.

Q50 R1: My [clippings of and writings on prior work] have website, link, title, or authors' names. If I can't [re-]find the work, I need to go through [web search] again. I wish everything is connected... once I click [on an element], it shows the [prior work file] and my [previously written down] ideas.

Q51 R1: [Collected prior work PDFs and my writings in physical notebooks] are not connected. [Having it] everywhere means you [have to] think over and over.

Q52 R3: Tying [writings] to PDFs is a big problem... Most of the time, when I take a [writing] out, I don't remember the context.

Many of our study participants call for new tools to address these breakdowns. Some call for visual environments, in which they can see curated works and prior writings, interact with them, and make new writings and other visual representations.

Q53 R6: Writing a review of literature section ... the points are: that you are aware of [a prior work], and how your work builds on it, is different from it, finds contrasting results to it, or tackles different problems. A lot of things you need to be aware of [and] keep in your head. I'd love tools that help us do that.

Q54 R11: I would like an electronic desktop that is as big as my real desk... If you have a big screen, you can see all the papers and zoom in-out much more easily.

Q55 R9: It would be interesting to [visualize] the space a set of references covers—How much overlap is there? What venues do they primarily pull from? Make sure I'm citing broadly, not only from one particular group.

3.3 An Interdisciplinary Model of Scholarly Research Ideation

We synthesize study findings, while incorporating concepts from information science, creative cognition, and art, to derive a new, interdisciplinary model of participants' scholarly research ideation practices and experiences. Our diagram of the new model (Figure 3) contains four components, corresponding to four categories of human action, in which investigators engage, while performing scholarly research ideation tasks.

Specifically, participants engage in *seeking* and *curating* prior work, and *reading* and *making sense of* found information, in order to *generate* new ideas and *explore* opportunities for further developing ideas. In practice, participants integrate and interweave these actions. Our diagram of the model shows this integration and interweaving with the interconnections among components.

We start by reviewing two dimensions—internal v.s. external, and learning v.s. creating—on which we arrange components in developing the model. The purpose is not to divide these components, but to show the integration of seemingly remote concepts and duality of scholarly research ideation. We then describe each of the four major components in the model, followed by elaborations on their interconnections.

Internal / External

We consider *external representation* as involving both physical and digital forms, perceivable in the environment, used to denote and organize concepts and thoughts in one's mind. Consistent with Hutchins, Scaife and Rogers, and the theory of distributed cognition, our study findings show that scholarly research ideation activities involve integration and interplays of internal / external representations and processes, which we will discuss in detail in this section. Our diagram of a new model of scholarly research ideation reflects this duality by arranging components in two rows, respectively representing internal / external actions, and shows the integration, by interconnecting components.

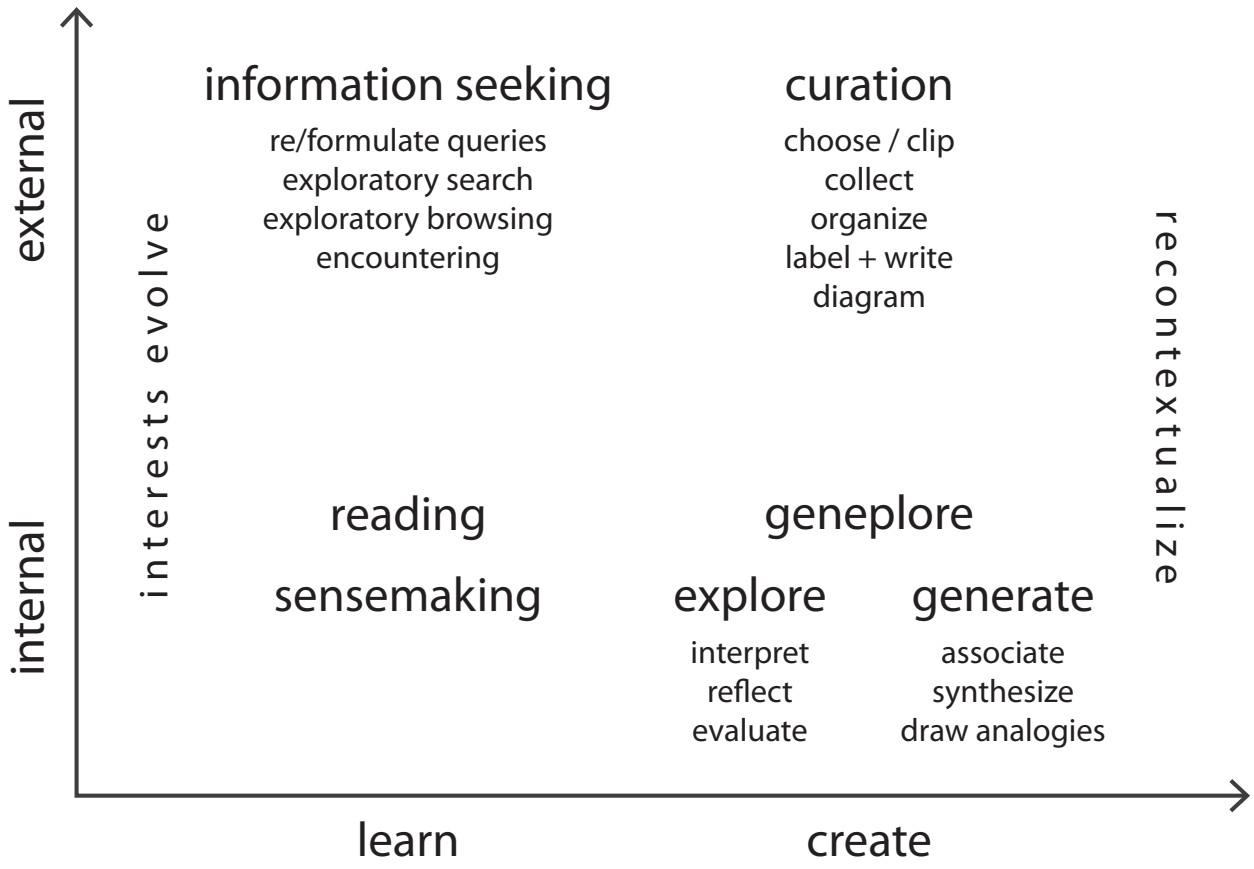

Figure 3: Ideation model integrates and extends existing models of creative cognition and information seeking, while assigning an essential role to acts of curation. Components, representing essential human processes involved in scholarly research ideation, are placed on two spectra: one corresponding to how much a process is internal or external, and the other corresponding to how much a process focuses on learning or creating. Curation, as both process and product, supports reading, sensemaking, further seeking, and ideation. "Evolving interests", as context, is placed above actions involved in seeking, and shown in italic.

Learning and Creating

Previous information seeking research largely focuses on people's processes of obtaining, understanding, and managing information relevant to their tasks. The goal is typically on acquiring knowledge, by developing representations that organize information and assign meanings, both in the mind and in the world. We characterize these processes as *learning* in general.

While learning processes are important, in scholarly research ideation tasks, researchers' goal

is *creating* new research products. To achieve this goal, they put together situated, evolving contexts of relevant information and understandings, and use the contexts to develop new ideas through Geneplore cycles. This perspective addresses the creative nature of doing scholarly research.

Processes of learning and creating are not mutually exclusive, but intimately connected. Learning often involves mini-c creative ideation, in which one develops ideas novel to herself. For example, prior study found that making annotations—a creative act—in reading facilitates readers to interpret contents and reflect on their circumstances [Marshall, 1997]. Knowledge acquired in learning often becomes foundations for developing new ideas. New ideas can in turn transform previous understandings, thus influencing learning experiences [Finke et al., 1992].

Spectra, rather than Exclusive Sets

While we differentiate the internal and the external, as well as processes focusing on learning and creating, we do not mean to draw definitive lines that demarcate exclusive sets of processes. Many processes involve both internal and external representations, and components of both learning and creating. For example, in sensemaking, people iteratively develop internal (mental) structures to encode external information [Russell et al., 1993]. This involves both "learning cycles" [Russell et al., 1993] and cycles for creating internal structures appropriate for the task at hand.

That said, to make studying and analyzing scholarly research ideation possible, we need to investigate different aspects of it. This is represented in our diagram (Figure 3) by components taking different positions on the two spectra: from internal to external, and from learning to creating. Each component represents an aspect of scholarly research ideation, and provides a perspective for narrower, deeper investigation. By placing processes on the two spectra, and contrasting and connecting them, we intend to reveal the inherent relationships among internal / external representations and processes, as well as actions of learning and creating. We are shifting the foci of processes, even as we understand that they overlap. In practice, researchers integrate and interweave these processes, in order to generate and develop new ideas for their work.

3.3.1 Component: Curation

We use an art-based concept of *curation* as a conceptual umbrella to denote current and prior findings of researchers' practices of developing contexts of relevant prior works, findings, and ideas. According to Kerne et al, strategies of curation include collecting, assembling, sketching, writing, and exhibiting [Kerne et al., 2017]. In our study, we observed uses of these strategies in participants' scholarly research ideation practices.

Collecting is the process of gathering content. Traditionally, collecting prior work has been a basis of doing scholarly research [Stoan, 1984]. Today, collecting information on the web, such as images, is a popular activity [Linder et al., 2014; Zhao and Lindley, 2014]. All participants collect interesting prior works. The strategy of collecting involves finding and choosing significant content. As Duchamp showed, through *Fountain*, choosing functions as a creative act [Lippard, 1971]. Clipping, the act of choosing a significant portion and saving a representation of it, is a situated method for collecting. Many participants clip and collect sentences, paragraphs, and figures from interesting prior works (R2's notes, R9's entries).

Assembling means fastening elements together to showcase the duality and tension between the original and resulting contexts. New meanings can emerge through assembling. In our study, many participants assemble. Types of elements they assemble include clippings from prior works, bibliography entries, hyperlinks, comments, and meeting notes. A particular form of assembling elements is *organizing* them, to form structure. In the present study, participants organize elements into folders (R2, R7), piles (R1), and maps (R3).

The *writing* strategy involves, at the macro level, developing arguments, and at the micro level, labeling and commenting on assembled content. Critical and reflective writings, in processes of consuming information, are found valuable for developing understandings and ideas [Marshall, 1997]. Labeling supports shifts of representations, in processes of sensemaking[Russell et al., 1993] and organizing information [Pirolli et al., 1996], helping practitioners identify categories, which become a basis of conceptualization [Lakoff and Johnson, 1999]. Social tagging, a collective form of labeling, is found vital for many online communities [Marlow et al., 2006]. Commenting

on found information online has become a popular means for creating new meanings. In our study, participants engage in both labeling and critical writing, to develop understandings and new ideas (Q33, Q43).

Through curation, researchers develop situated contexts that relate diverse information, collected from environments they live in and work with, including interesting prior works and researchers' own thoughts. Curated information is externally represented by clippings and other element forms. Researchers see and interact with these external representations, which they collect and organize in situated curation contexts. This supports their (embodied) cognitive processes: e.g., developing understandings, discovering relationships, and having new ideas. When new information is curated, researchers see and think of it in the context of other assembled information. This *recontextualization* provides a basis for new understandings. Our study finds that participants engage in curation throughout their scholarly research ideation activities.

Our conceptualization of curation focuses on its generative connotations—through curation, researchers develop new, situated contexts of diverse elements. In these situated contexts, researchers can see and interact with external representations of information and ideas, which in turn supports their (embodied) internal processes: developing understandings, discovering relationships, and having new ideas. Newly curated elements are perceived in the situated context, promoting new interpretations and discoveries of a multiplicity of relationships. Our notion of curation complements those used in information science and e-science (e.g., [Abbott, 2008; Gray et al., 2002; Higgins, 2008; Whittaker, 2011]), which primarily focus on transactional aspects of collecting, managing, and archiving information. To archive means to put content away, while to curate means to exhibit it. Through exhibition, people experience collections. Thus, curation in art and scholarly research ideation serves to actively recontextualize content.

Role in the model. *The role of curation, in the model as a whole, is to provide a process and space for iteratively creating a new, task-specific conceptual context, building on and with found prior work, thus to (1)* ***recontextualize*** *found prior work, and (2) support developing understandings and ideas.*

In art, curators actively (re)contextualize the significance and meanings of found artworks, to conceptualize and frame new meanings [O'Neill, 2012]. Similarly, in scholarly research ideation, through curating found prior works, researchers *recontextualize* their significance and meanings. Conceptually, this process resembles Ernst clipping images out of medical catalogues, then recontextualizing and transforming them, in developing Dada collages [Krauss, 1994].

For example, Page et al recontextualized the idea of using citation indexes to measure impact of peer reviewed scholarly articles [Garfield, 1955, 1972], out of its original, scientometric field, bringing it into the (then) new context of web search [Page et al., 1999]. The result was an impactful algorithm, PageRank, which created the basis for one of the world's most successful companies, Google. In the PageRank paper, the authors first reviewed found prior works within their original contexts, such as academic citation analysis [Garfield, 1995]. They then discussed the differences between the original, scientometric context and the new, web context. They explained how they improved prior methods to address the challenges of the new context. This differentiation from prior work advanced the initially recontextualized idea, becoming the basis for Pro-C research and Big-C societal contributions.

In our study, many participants, such as R9, curate reflective writings with interesting prior work, and revisit them when writing new papers. R3 has a bag of "snippets" and "quotes" for use in her writings. R13 curates writings in the form of "cuts" and "notes". These acts of curation support participants in iteratively developing understandings and ideas, through cycles of representing them in text and editing. Curation thus supports integration of internal and external representations, and interplays of internal and external processes, providing a basis for reflection-in-curation [Webb et al., 2013]—in which participants look back and think at what they have collected so far—which iteratively stimulates further scholarly research ideation.

3.3.2 Component: Geneplore

The new model of scholarly research ideation incorporates the Geneplore model of creativity, which connects generative and exploratory cognitive processes (emphasizing internal) [Finke et al., 1992]. In generative processes, one develops preinventive structures, i.e., internal representations

possessing properties promoting creative discovery. In exploratory cognitive processes, one pursues opportunities to develop meaningful and novel ideas using preinventive structures, or, in the case of failing to do so, modifies preinventive structures through focusing and expanding concept, in preparation for another Geneplore iteration. In our study, many participants engage in *synthesis* and *forming analogies* to develop preinventive structures, often using curated content such as prior work, and use *interpretation* and *reflection* to further develop new ideas for scholarly research.

Synthesis, the act of combining multiple elements into a new, meaningful whole, is an example of generative processes researchers engage in. One of R2's most cited papers is a synthesis of prior concepts, related to an emerging field. R6 synthesizes prior ideas in new ways to make new contributions (Q5). R9's seminal work on a familiar task in an emerging setting synthesizes collected data into interesting findings, which have motivated ongoing development of her field.

Writing research proposals and papers involves synthesis. Authors need to combine knowledge, experiences, insights, thoughts, and study/experiment results into coherent, valuable, and novel arguments (Q53), while developing a new research perspective and framing. This is often an iterative process, in which one experiments with multiple ways of combining elements, developing relationships, and deriving meanings.

Many participants *form analogy* across domains to help develop new ideas (Q22 , Q6 , Q8). To form an analogy, a researcher needs to find another domain, in which the *relationships* among entities are structurally similar to those in the current domain [Gentner, 1983]. Then, the researcher can draw ideas, theories, and methods from the other domain, to help understand phenomena and solve problems in the current domain.

Interpretation is an exploratory process, through which people contextualize and derive new understandings. When engaging with prior work, a researcher needs to interpret, to understand how the work is positioned in the field and how it relates to her own research (Q25). Interpretation of prior concepts can lead to discovery of new perspectives and findings (Q11 , Q23). Interpretation is rooted in perception [Merleau-Ponty, 1962]. Thus, forms of external representations, such as clippings, mind maps, sketches, and drawings, and the ways external representations are put

together and combined, affect perception, hence interpretation.

Reflection is another exploratory process, in which people look back on their understandings and practices, to derive and articulate previously implicit knowledge [Schön, 1992]. Reflecting on one's own research while reading related work helps researchers articulate their perspectives and develop new ideas (Q22).

Role in the model. *Geneplore draws on cognitive psychology to enable understanding people's internal creative processes of generating and developing ideas.* Scholarly research ideation, as the task, motivates constituent processes. Seeking, reading about, and making sense of relevant prior work serves the purpose of achieving ideation, e.g., by supporting researchers in gaining understandings, articulating perspectives, deriving contributions, and stimulating inspirations. However, as participants pointed out, seeing prior work in an early stage of research can potentially cause fixation. Curation supports ideation, as external cognition, by supporting relating and recontextualizing found information, and further promoting Geneplore discovery and exploration of unexpected, situated relationships leading to new ideas.

3.3.3 Component: Seeking

This component involves actions researchers take to find interesting information, in order to change their states of knowledge about the topics they are working on [Marchionini, 1997] (Q21 , Q13 , Q14). Seeking is motivated by people's *interests*, which we characterize as their situated and spontaneous desires for particular information. An interest can be implicit, i.e., unarticulated or unrecognized, even as it serves as a basis for action. Encountering information related to implicit interests can produce experiences of serendipity [Blandford and Attfield, 2010; Foster and Ford, 2003] (Q4). An explicitly recognized interest resembles Belkin's anomalous state of knowledge [Belkin et al., 1982].

Through information seeking, people pursue their interests, to resolve their anomalous states of knowledge. Search and browsing are two complementary methods [Marchionini and Shneiderman, 1988]. In search, people formulate queries to represent their needs, and use a computational system, such as a search engine, to find documents relevant to their queries. In browsing, people

opportunistically follow promising leads, trying to find new information. Whimsy and serendipity are in effect. Articulating needs, formulating queries, and picking leads to follow are often (mini-c) creative ideation processes, in which people generate and develop ideas of what specific information can serve their interests, and how they can find it.

Like ideation processes, seeking is situated in contexts. Personal experiences, domain knowledge, ongoing activities, and social relationships are all constituents of context, which affect how one searches and browses. Contexts are, at the same time, outcomes of seeking processes. For example, in our study, when seeking prior work in unfamiliar fields, many participants start with general queries or known work, find new work through search and browsing, and pick up terms and keywords in newly found work (Q21). In this process, their knowledge, which serves as a personal context for seeking, changes. Their interests evolve, becoming more specific. Phenomena of *evolving interests* are a basis for exploratory search [Marchionini, 2006; White et al., 2006a; White and Roth, 2009; Wilson et al., 2010] and browsing [Bates, 1989; Jain et al., 2015; Qu et al., 2014].

Role in the model. *Information found in seeking, including prior work of interest, contributes to contexts in which researchers develop ideas and derive novel contributions.* In general, good scholarly research builds upon related prior work, yet differentiates itself in meaningful and novel ways. Seeking relevant prior work is regarded as an essential skill of practicing scholarly research in the time-honored notion of "doing research in library" [Stoan, 1984]. Hyland recognized that, in scholarly writing, appropriate citations to related prior work display allegiance to a target community, help establish a framing for the author to derive new contributions, and thus are valuable for "establishing a persuasive epistemological and social framework for the acceptance of [the author's] arguments" [Hyland, 1999]. In practice, in using found prior work to support claims of valuable and novel research contributions, researchers further engage in reading, sensemaking, curation, and Geneplore processes.

3.3.4 Component: Reading and Sensemaking

Researchers read found prior works, to make sense of them, and understand how it can be useful to their own work. Participants' reading processes are combined with critical thinking, through

acts of underlining, highlighting, and scribbling comments, thus constitute *active reading* [Adler and van Doren, 1972]. In sensemaking, a researcher engages in shifting representations of found information, to make them appropriate to her scholarly research ideation task at hand [Russell et al., 1993]. Through active reading and sensemaking, researchers develop understandings of found prior work, people, and relationships (Q28), which are organized in knowledge structures, such as mental models and schemas. *Mental models* are internal simulations of how things in the world relate to each other and work together [Gentner and Gentner, 1982]. *Schemas* strcturally organize information about objects and relationships [Brewer, 1987].

In many cases, domain knowledge is necessary or beneficial for developing new ideas [Finke et al., 1992]. Thus, active reading and making sensemaking can be important for scholarly research ideation. Many innovations build upon understandings on and insights in precedents. Appropriate external representations facilitate active reading and sensemaking (Q34).

Role in the model. *Reading and sensemaking support, and sometimes are integrated with, other actions in scholarly research ideation.* Previous information science research has found that sensemaking and learning are essential for information seeking. As the seeker makes sense of and learns about the topics or fields being investigated, their interests and needs evolve, leading to changes in how they seek information (e.g., reformulation of queries). Making sense of curated elements—in the curation context—is necessary for developing situated understandings and discovering unexpected relationships. The outcomes of reading and sensemaking, e.g., mental models and schemas about interesting prior work, become stimuli and building blocks for new ideas.

3.3.5 Interconnected Components

By assembling concepts from diverse disciplines, this interdisciplinary model of scholarly research ideation presents interconnections among actions represented by its components, some of which are previously understudied. The interconnections—between internal and external, learning and creating—are essential. These processes are inextricably intermeshed. We elaborate on these interconnections.

Curation supports reading, sensemaking, and Geneplore processes. Through curation, researchers develop a situated, evolving context for their scholarly research ideation tasks at hand, supporting recontextualization of found prior work, information, and ideas in it. This recontextualization is essential for actively reading about and making sense of how diverse prior works relate to each other, and how they connect to one's own research. For example, researchers write about found information as they seek and curate prior work. Through writing about found prior work, an act of curation, they externalize their understandings, with a focus on how it compares to or becomes useful in their own work (Q39 , Q40). These acts of writing thus function like note-taking in reading, which has been found to facilitate learning [Kiewra, 1989], while potentially going further, to manifest ideation.

Recontextualization is also essential to many Geneplore processes, such as association, synthesis, interpretation, and forming analogies, since these processes relate diverse elements. In our study, participants put together curations to support Geneplore processes. For example, when writing research papers, some participants revisit their curations, in order to think across prior work and develop new perspectives and framings (Q45 , Q25).

By supporting bringing material from diverse contexts together, curation addresses the situatedness of reading, sensemaking and ideation processes. By enabling researchers to choose and clip content, as well as organize, label, tag, and write about collected elements, curation helps researchers integrate internal and external representations, which is found crucial for supporting and enhancing cognition [Kirsh, 2010; Scaife and Rogers, 1996]. Thus, it is no surprise to us that all participants engage in curation of prior work in their scholarly research ideation activities: curation addresses their needs to understand, relate, and create.

Curation and Geneplore cycles are integrated. The process of curation involves iterative cycles of reflection and interpretation, as the curator explores how collected elements relate to each other. New ideas can emerge from these exploratory cycles. Through organizing, labeling, and writing, the curator generates and develops a sense of how to articulate discovered relationships and emerging ideas. These manipulations serve as external cognition for internal creative

processes. New understandings of relationships and new ideas in turn help the curator decide what to seek, choose, and curate, constituting reflection-in-curation [Webb et al., 2013]. Thus, curation and Geneplore cycles are integrated. R13's use of "notes" and "cuts" is an example of iterative ideation in conjunction with curation. This is consistent with prior findings on people integrating internal and external representations and processes in performing cognitive tasks [Scaife and Rogers, 1996].

Geneplore processes manifest in seeking, reading, and sensemaking. Geneplore processes of creative ideation manifest in many actions involved in scholarly research ideation activities, including seeking, reading, sensemaking, and curation.

In information seeking, the seeker starts with interests (e.g., feeling compelled to investigate a research problem), recognizes anomalous states of knowledge (e.g., lack of knowledge of related topics), develops situated information needs (e.g., "I need to find some good and recent review articles on this topic, using whatever digital libraries the university provides"), and formulates queries, to address their needs, resolve their anomalous states of knowledge, and, eventually, serve their interests. This is a mini-c creative process. It involves Geneplore processes, such as reflection (on what information is needed) and synthesis (of queries). In some other cases, Geneplore cycles reveal anomalous states of knowledge and needs, e.g., when the researcher finds a missing piece in the process of synthesis, or reflects on her understandings and practices. This is seen in study findings (Q18 , Q20 , and Q21).

In sensemaking with found prior work, researchers search for representations that best support the scholarly research ideation task at hand. This search involves creative, Geneplore processes, including initial generation of ideas on how to represent information, evaluating representations, and potentially transforming (shifting) representations. Representations afford interpretation. Novel interpretations potentially lead to new insights [Finke et al., 1992]. In our study, R7 mentioned a case where new interpretations of an existing concept make valuable contributions.

Seeking, reading, and sensemaking support Geneplore cycles. Internal and external cognitive processes are inherently inseparable. Internal representations developed in reading and sense-

making, with information found in seeking, are important for subsequent Geneplore cycles. These internal representations include mental models and schemas. Mental models can be blended to form preinventive structures [Finke et al., 1992]. Schemas can support exploration of novel interpretations and imaginations, by suggesting relationships between concepts [Finke et al., 1992]. Previous empirical studies show that external information has the potential to stimulate new ideas [Shah, 1998]. On the other hand, existing mental models and schemas can also cause fixation, preventing new ideas from generation (Q9 , Q10) [Finke et al., 1992]. In our study, R3 flips through curated papers to trigger ideas and potentially overcome fixation in writing (Q45).

Seeking is integrated with reading and sensemaking. Previous information science research found that information seeking is inherently integrated with gaining knowledge (e.g., through active reading) and sensemaking [Blandford and Attfield, 2010; Marchionini, 1997, 2006; White et al., 2006a; White and Roth, 2009]. People read about and make sense of found information, to develop understandings of unfamiliar topics and fields. New understandings in turn change how people seek, make sense of, and learn about information. This is seen in participants' practices (Q2).

Participants' active reading practices include acts of underlining, highlighting, and scribbling comments. These acts are also acts of curation. Comments scribbled during active reading are often interpretive [Marshall, 1997]. New ideas can emerge in this act of writing interpretive comments.

Curation, seeking, reading, and sensemaking are interwoven. Curation provides a means for representing researchers' evolving interests, while they develop understandings of collected prior work, and develop new ideas. Because curation supports making sense of found information in the evolving context, of the ongoing scholarly research ideation task, researchers interweave curation and seeking when investigating unfamiliar topics or fields. They thus develop situated understandings while seeking and finding interesting work. The amount of found information could be large, leading to needs for choosing and clipping. Curation also provides means for exploring relationships, through collection and organization, and articulating intermediate findings and emerging ideas, through labeling and writing. New understandings and ideas developed in the

curation context can in turn lead to new needs, and reveal new directions, motivating further seeking. The popular use of transient in-browser curation in participants' practices, in which people curate found information using browsers and tabs, is an example of how seeking becomes interwoven with curation. Limitations in support for transient in-browser curation in practice lead to breakdowns, such as struggling to see complex relationships.

3.4 Implications for Design

Through the model of scholarly research ideation, our findings provide a basis for deriving implications for the design of investigations of and support for scholarly research. Researchers interweave and integrate various actions in performing scholarly research ideation tasks. Thus, in the design of studies and interfaces addressing researchers' practices, we advocate addressing scholarly research ideation tasks as a whole, rather than atomized actions, such as search and sensemaking. In particular, we advocate integrated research to investigate and support how people collect, assemble, and write about prior work, which, drawing on art practice, is known as *curation* [Kerne et al., 2017].

Curation, as such, is deeply interwoven with scholarly research ideation processes. We prescribe that curation in scholarly research ideation would benefit from supporting users in collecting elements of diverse material and assembling them with flexible media. Transient in-browser curation is a particular form. Our methodology—of drawing from art as a means for developing theories of how to support creative processes—has the potential to be valuable in future studies of ideation across fields.

3.4.1 Contextualizing with Prior Work is Essential for Ideation

All participants agree that familiarizing with and properly citing prior works is important. Here, prior works include not only research done on exactly the same topic, but also those on related and (seemingly) unrelated topics.

Prior works constitute a context, which researchers build upon, draw from, and compare to, to argue for new contributions. For participants working in various discursive fields, such as art and

humanities, referring to prior works is an inherent part of creating new work (e.g., Manet's *Olympia* which alludes to Titian's *Venus of Urbino* [Clark, 2015]). In fields that we think of as more linear, such as biology and physics, while many new ideas come from data generated by experiments and studies, *making valuable interpretations of data often requires knowledge of prior work* (e.g., R4 making interpretations of a new pattern revealed in data analysis). Thus, contextualizing one's work with prior works seems to be an essential part of scholarly research ideation. Tools addressing scholarly research ideation need to support such processes of contextualizing with prior works. Future research should also investigate how data is examined in and becomes part of this context.

3.4.2 Integrate Internal / External to Support Ideation

The integration of internal / external representations and processes is essential for many human activities. In sensemaking, people integrate processes of developing internal representations, such as schemas, and manipulating external representations, such as collections and visualizations [Russell et al., 1993]. Designers often engage in conversations with external representations, such as a "big wall" of materials, in generating new design ideas [Kolko, 2010; Schön, 1992]. Artists sketch, not only to record ideas, but also to help generate new ideas, through perceiving and recognizing [Goldschmidt, 1991].

Therefore, this integration of internal / external representations and processes should be addressed by investigations of scholarly research ideation tasks, and more generally, tasks with a sensemaking or creative constituent. Interfaces supporting these tasks need to address this integration in practice. This principle is the basis for developing the role of curation in these tasks, which we discuss in the next subsection.

3.4.3 Support Curation to Aid Ideation

Study findings show that curation plays a crucial role—to facilitate contextualization—in scholarly research ideation. Contextualization is essential for creative ideation. Participants engage in prior work curation throughout their scholarly research ideation activities. Through curation, researchers put together contexts, in which they integrate external/internal representations and pro-

cesses, to make sense of information, discover relationships, connect, synthesize, compare, contrast, and, ultimately, develop new ideas. This role of curation is exemplified by participants' use of transient in-browser curations in seeking, and R13's use of "notes" and "cuts" in writing. We now draw from [Kerne et al., 2014] to discuss how to better support curation in scholarly research ideation.

3.4.3.1 Elements of Curation: Supporting Diverse Materials

In architecture, tectonics involves processes of joining, which are based in the properties of materials [Webb et al., 2016]. Each space for human activity, as a whole, derives from these properties and how joining is articulated. Thus, according to a tectonic theory, curation involves two levels: the materials of elements and media of curation. The former determines how individual elements can be collected, perceived, interpreted, and interacted with.

Our study participants use diverse materials of elements in their prior work curations, e.g., PDFs, bibliographic information, textual notes, and browser bookmarks (see Section 3.2.5.1). Each material has its advantages and disadvantages. For example, R9's writings (Q39 , Q40 , Q41), containing citations and her thoughts, are easy to make, yet unable to capture visual information, such as figures and diagrams. In our study, many participants combine elements of multiple materials in their prior work curations.

We recommend supporting diverse materials of elements, to address researchers' diverse and situated needs. For example, we recommend supporting both images and text clippings. Image clippings, including figures and thumbnails, are effective in representing visual and non-verbal content, supporting embodied cognitive processes. Textual writings are useful for representing conceptual categories and ideas. Combining image and text representations is found to facilitate formation of mental models [Glenberg and Langston, 1992], thus promoting understanding and ideation.

3.4.3.2 Media of Curation: Supporting Flexible Assemblage

Media of assemblage define how elements can be joined together to form a whole, thus affecting how people can interpret relationships among elements, and discover new ones. Common media of assemblage used by participants are relatively *rigid*, i.e., they make it difficult to represent complex, situated relationships and schemas among curated elements.

For example, many participants curate PDFs of interesting articles in a file system's hierarchy of folders (e.g. R3, R7). This medium of assemblage imposes *premature formalism* [Shipman and Marshall, 1999], in two aspects:

1. With file system hierarchies, it is common that one element—a file or folder—can only appear in one containing folder at a time. However, researchers' understandings of how to categorize curated elements are often ambiguous (Q32 , Q33). In many cases, the boundaries among categories are fuzzy, and a single element spreads multiple categories. File system hierarchies thus fail to support researchers' fuzzy categories.

2. Researchers' categories evolve (Q31). Yet, changing a file system hierarchy usually involves considerable effort. Previous research stated: "Formalisms are often difficult for people to use because they need to take many extra steps (and make additional decisions) to specify anything." [Shipman and Marshall, 1999] File system hierarchies, a kind of formalism, thus fail to address researchers' evolving understandings.

Linear list is another common medium of assemblage (e.g., R2's Evernote collection, Figure 1). Human textuality, such as citation networks of scholarly publications, is often nonlinear [Landow, 2006]; thus, it cannot be directly represented in linear form.

In contrast, spatial information organization is a method for reading and writing with physical paper [Sellen and Harper, 2003]. This method can be informal and flexible, supporting fuzzy categories, evolving understandings, and non-linear relationships. Arranging elements in space can create a multiplicity of relationships, stimulating interpretations and discoveries. Thus, spatially organizing information is found to promote development of new ideas [Kolko, 2007; Marshall

and Shipman, 1995]. In our study, some participants use paper printouts as a flexible medium for assembling relevant prior work (Q33). Others call for flexible digital media for spatially organizing their prior work curations, to help them see relationships (Q47 , Q54 , and Q55). We recommend that future research design and evaluate new interfaces providing informal and flexible curation spaces.

3.4.4 Supporting Associating Ideas and Making Analogies

Associating ideas and making analogies across domains are popular methods participants use to conceptualize novel research. Associating (dissimilar or distant) ideas develops a multiplicity of relationships, and promotes emergence [Wilkenfeld and Ward, 2001]. Linsey et al use analogy as a tool for producing novel engineering design [Linsey et al., 2008]. Forming analogies requires understanding relationships in one domain, and finding another domain with structurally similar relationships.

Thus, tools addressing scholarly research ideation should support associating ideas and making analogies across domains. One way to do so is to facilitate free-form juxtaposition, arrangement, and composition of diverse external representations of information and ideas, promoting users to make novel interpretations, abstract, and discover hidden relationships necessary for forming meaningful associations and analogies. Support for sketching and annotation can be important. Sketching is an embodied physical process of visual thinking, which represents and develops relationships [Kerne et al., 2017]. Annotation is a form of textual writing, which can be used to develop conceptual categories. Prior research shows that combining textual and diagrammatic representations help people develop mental models [Glenberg and Langston, 1992].

4. A DYNAMIC EXPLORATORY BROWSING INTERFACE FOR CITATION CHAINING *

Findings from the study with established investigators and the derived model of scholarly research ideation suggest guidelines for the design of interfaces supporting researchers' work:

1. Support curation of prior work in the emergent context of an ideation task. Curating, making sense of, and reflecting on prior work is expected to develop the curator's thinking.

2. Support integration of internal and external representations, to promote ideation.

3. Support mini-c creativity. Personal learning and intermediate findings are stepping stones that lead to discoveries of larger impact.

With these guidelines in mind, we designed and developed a new interface, the Metadata In-Context Explorer (MICE). MICE addresses a fundamental activity involved in scholarly research ideation—browsing. According to Marchionini and Shneiderman, "browsing is an exploratory, information-seeking strategy that depends on serendipity" [Marchionini and Shneiderman, 1988]. By *exploratory browsing*, we mean browsing when the task is open-ended and the user is unfamiliar with the space of information. *Citation chaining*, the process of recursively finding works citing the current work, to understand how the original work is commented on, followed up, and used, is a form of exploratory browsing. Exploratory browsing is key to *berrypicking* [Bates, 1989], the iterative process in which the user encounters new information, and her understanding and information needs evolve. Browsing and search are complementary strategies for exploring information [Marchionini and Shneiderman, 1988]. In *exploratory search* [White et al., 2006b], users engaged in learning and investigation iteratively refine information needs. While this chapter directly addresses exploratory browsing, its implications also impact exploratory search.

*Edited reprint with permission from "Metadata Type System: Integrate Presentation, Data Models and Extraction to Enable Exploratory Browsing Interfaces" by Yin Qu, Andruid Kerne, Nic Lupfer, Rhema Linder, and Ajit Jain, 2014. In *Proceedings of the SIGCHI Symposium on Engineering Interactive Computing Systems*, pages 107-116, DOI: `https://doi.org/10.1145/2607023.2607030`. Copyright 2014 by Qu, Kerne, Lupfer, Linder, and Jain. Publication rights licensed to ACM.

Researchers engage in exploratory search and browsing of prior work in scholarly research ideation (see Section 3.2.3). In a typical scenario, researchers search and browse relevant prior work, berrypick and curate interesting findings, and revisit when conceptualizing a research project or writing a research paper.

People encounter issues when performing exploratory browsing tasks with existing interfaces and tools. People may lose context while browsing, as new information is encountered [Edwards and Hardman, 1999]. Interlinked pages are shown in separate viewports, leading to *disorientation* [Conklin, 1987], the problem of not knowing where you are or how to return to an encountered page in a network of information, and *digression* [Foss, 1989], the problem of going off track amidst many open windows or tabs. Disorientation and digression can grow acute during citation chaining, and further impedes scholarly research ideation. To counter, we design new interfaces that maintain context for the user during exploratory browsing, out of which MICE is an example.

Exploratory browsing interfaces that maintain context requires dynamically presentation of trails [Bush, 1945] of linked documents. Since display and the user's cognitive resources, such as working memory [Cowan, 2001], are limited, the interface must present documents as summaries, to reduce display space and cognitive load. Typically, summaries are manually or algorithmically extracted texts that reveal the chief points or substance of a document.

In this thesis, we use *metadata* or *semantics* to refer to contextual information describing a real world entity in detail, as well as its computation representations affording broad user actions, such as reading, collecting, searching, and sharing. For example, information about a product—including its name, price, specifications, pictures, and reviews—is considered product metadata. *We hypothesize that metadata—when properly presented such as in MICE—can function as a valuable form of summary for supporting exploratory browsing.*

Many popular, useful web sites do not directly publish metadata. They instead present metadata in ways specifically designed for human readers, often with other information, such as navigational guides and ads. Thus, to present metadata in our interface, mechanisms for extracting metadata from ordinary web pages are required. Further, exploratory browsing involves encountering het-

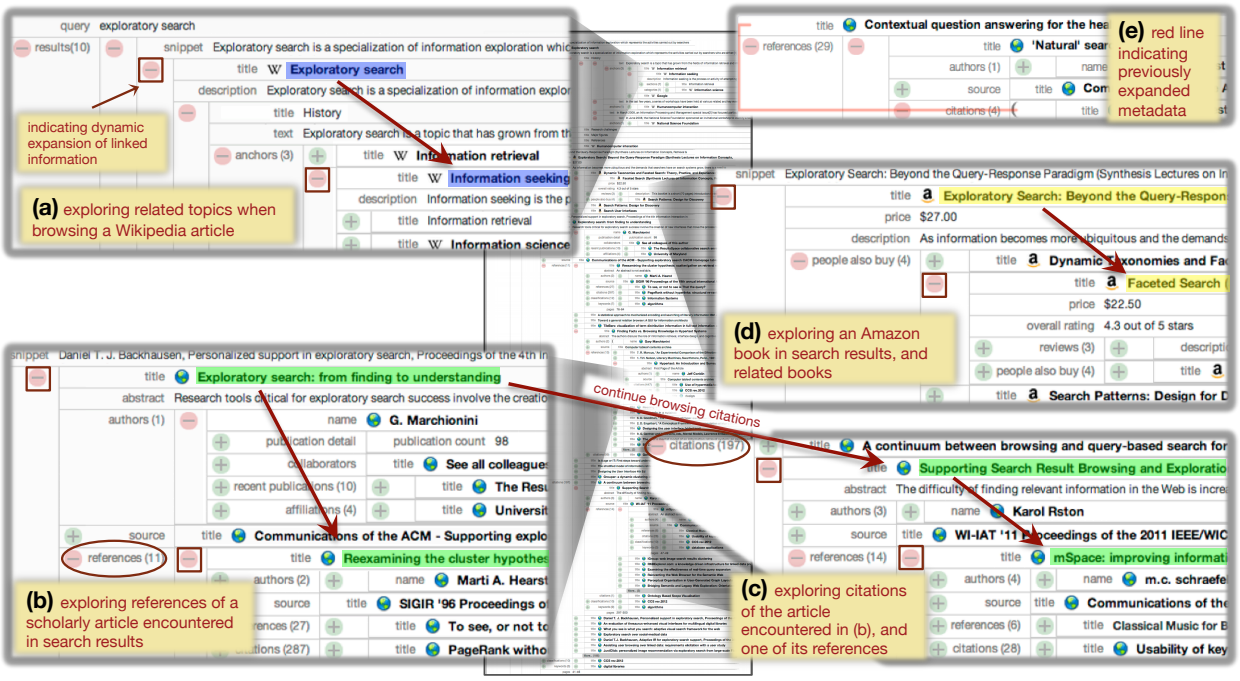

Figure 4: An overview of Yin's exploratory browsing with MICE. Snippets show close-up views of her session. Arrows denote browsing linked information.

erogeneous types of information, each of which is best represented with particular data models, layouts, and styles. The interface must dynamically derive and present metadata summaries particular to encountered information types at runtime, to reduce the cognitive load of browsing chains of linked information.

We operationalize a metadata type system to address challenges of dynamically summarizing and presenting heterogeneous documents as expandable metadata in a unified context, to enable exploratory browsing across diverse websites. We will discuss the details of this metadata type system in the next chapter.

Using the metadata type system, we built an *example* dynamic exploratory browsing interface, the Metadata In-Context Expander (MICE) (Figure 4) [Interface Ecology Lab, 2014]. The visual appearance of MICE looks like a typical XML or RDF visualizer, but goes beyond by supporting dynamic acquisition, presentation, and exploration of new information. MICE makes relationships

between linked documents visible.

We begin this chapter with a brief introduction to the interface of MICE, followed by an exploratory browsing scenario that motivates development of the interface. We then present a user study in which computer science students browse and explore scholarly articles for prior work search and project ideation. The results show how MICE supports exploratory browsing in context. Based on the study, we derive implications for engineering dynamic exploratory browsing interfaces to support scholarly research ideation tasks.

4.1 The MICE Interface

MICE is a general dynamic exploratory browsing interface for presenting complex, dynamic, interlinked metadata. This section provides an overview of the MICE interface (Figure 4).

MICE presents a metadata record as a list of fields. A field can be as simple as a title or a price, and as complicated as a list of search results. Each field is presented in two parts: name and content. The name indicates the purpose of this field, whereas the content fulfills this purpose. Name is presented in smaller font and lighter colors, to create a layer, and to feature content. When the content is long (e.g. a paragraph), the initial presentation will truncate it, but hovering over will reveal the complete content. When the field has complex internal structures, e.g. when it is another metadata record, or a list of multiple records, it can be expanded.

An important feature MICE supports is the dynamic expansion of new metadata records. When expanding a field that points to an external document, MICE will dynamically access the document, figure out which metadata type is best for it, extract semantics from the document using the best metadata type, and presents it according to the presentation hints specified in the metadata type. The new metadata record will be presented in the same context as the source record, and their relationship will be made clear by the field name (Figure 4a, b, c, d).

When browsing interlinked metadata records, it is possible to encounter one that is already present in the context. In that case, MICE will display a red line to indicate its first appearance in the current context (Figure 4e).

4.2 Scenario

Yin is a computer science student who wants to conceptualize a research project about exploratory search. He starts by searching Google for "exploratory search". MICE presents search results in an expandable list. Each result is presented as a snippet followed by a collapsed document with only the title visible. Yin notices that the title shown in MICE is clickable, like the title links Google would present. He clicks on the plus button to expand the collapsed document in the first result, which is a Wikipedia article [Wikipedia editors, 2018]. Information from that document is accessed in real time, converted to metadata summary, and presented in MICE (Figure 4a).

The metadata summary of the Wikipedia article introduces the concept of exploratory search, including its history, research challenges, and major researchers. Related topics are linked. Some linked topics are new to Yin. With MICE, Yin can easily expand linked topics, bringing metadata about them into the current context (Figure 4a). Further related topics are again linked as expandable metadata. Using this recursive information expansion, he explores topics, such as "information seeking", "faceted search", and "information foraging". He can still see the central topic, exploratory search, at the top, which helps her maintain focus without being disoriented by Wikipedia's many links. After a period of exploration in Wikipedia, he collapses these related topics in MICE and returns to the search results.

Wikipedia provides a good overview, but Yin wants to dig deeper. From the search results, he expands a scholarly article [Marchionini, 2006] from the ACM Digital Library. MICE extracts metadata for that article, including authors, abstract, references, and citations, and presents it in a concise form (Figure 4b).

Yin expands that article's references in MICE, seeking prior work it builds upon. He sees an interesting article [Hearst and Pedersen, 1996] (Figure 4b) about exploring information, in which the user starts with a broad query and uses clustering to gradually refine the scope of information to explore [Cutting et al., 1992]. He finds this idea inspiring. Another reference [Marchionini and Shneiderman, 1988] distinguishes two different strategies for finding information: search and

browsing. Yin keeps chaining references, discovering a seminal survey on experiential problems in hypertext [Conklin, 1987], e.g., the cognitive load of seeing enormous amounts of information and maintaining context. By exploring prior work, Yin expands her understanding of the roots of research in this field.

With this understanding, Yin seeks new work in this field. He goes back to a scholarly article encountered in the search results, and expands its citations (Figure 4c). MICE shows 10 citations at a time, out of 197. Yin clicks on a button at the end of the list to reveal the next 10. He expands a title that catches her eye: [Rástočný et al., 2011] (Figure 4c). It discusses an interesting idea of applying zoomable user interfaces to clustering-based exploration. From the references of this article, he notices another one [schraefel et al., 2003] by one of the major researchers introduced in the Wikipedia article. He expands it (Figure 4c).

As Yin continues exploring, he encounters information from sources including CiteSeerX, Google Books, and Amazon (on which he orders books on exploratory search and faceted search, Figure 4d). MICE supports her exploration process by showing rich, useful metadata summaries of browsed documents, and iteratively bringing linked information into context on click. When he encounters a previously expanded document, MICE displays a red line on hover, leading her to the previous occurrence (Figure 4e). Over an extended period of non-linear exploratory browsing, Yin learns a lot about the topic, including motivation, prior work, critiques, and new directions. Synthesizing what he learns, Yin conceptualizes a project about software architecture that supports multiple paradigms of exploratory browsing and search interfaces.

4.3 User Study

We designed and conducted a 2x2 within-subjects experiment to validate our hypotheses that: (1) metadata will serve as an effective form of summary, and (2) dynamic exploratory browsing interfaces like MICE will support exploratory browsing tasks better than a typical web browser. In the task context, students from an information retrieval class used *citation chaining*, the process of following references, citations, and authors for exploratory browsing, to conceptualize a project for the class. Citation chaining is a common task researchers perform for scholarly research ideation.

Independent variables we manipulated were *initial document set* and *interface*, each with two conditions. The instructor picked two topics for initial document set: *query log* and *PageRank*. Each initial set consisted of 7 scholarly articles from ACM Digital Library, IEEE, or CiteSeerX. The experiment interface condition uses MICE for exploratory browsing, while the control interface condition uses a regular web browser and hyperlinks.

We recruited 8 undergraduate (1 female, 7 male) and 5 graduate (all male) students who were taking or had taken the class. 12 of them have used scholarly digital libraries and repositories, such as ACM Digital Library and IEEE Xplore, to find prior work. None of them, nor the instructor of the class, was affiliated with our lab. The study process for each participant consisted of a survey (5 min), an introductory video (5 min), two sessions of exploratory browsing (25 min x2) with different initial document set and interface conditions, and a survey (5 min). Conditions were counterbalanced. In each session, the participant spent 5 min on an interface tutorial video before engaging in exploratory browsing with papers. Participants used CiteULike to collect interesting papers in all conditions.

We recorded browser interactions and collected articles. A two way ANOVA shows students spent significantly less time directly browsing digital library web pages when using the MICE interface: .83 minutes compared to 16.43 minutes for the control ($p < .001$). This indicates that though the students could browse the original digital library web pages from MICE, they overwhelmingly did not need to, since metadata summaries presented by MICE were sufficient for them to perform the task. There was no significant difference in the number of collected papers between conditions.

The questionnaire asked participants about their experiences with both interfaces, gathering Likert scale quantitative and open-ended qualitative data. The Likert scale ranges from -4 (strongly preferring control interface) to 4 (strongly preferring MICE). Participants rated MICE better than the control in all four dimensions of experience. A single sample one-tailed t-test with α = .95 and $\mu < 0$ as the alternative hypothesis showed statistical significance for each (Table 2).

Qualitative data analysis depicts aspects of user experience:

1) *Concise & clear representation.* Users reported the concise representation of metadata summaries helped them browse while citation chaining:

> u8: [MICE] provides a much better method to chain documents by saving space and condensing the data for users to read and skim through.
>
> u12: The compactness of the UI makes it easier to go through a chain without losing track of where you started.

Participants value the way MICE clearly organizes information with categories, such as references and citations.

> u13: Categorization, i.e., (of) abstract, source, references, (helps citation chaining). (Information is) organized well.
>
> u12: Information about each paper is organized in a clear fashion.

2) *Less digression.* Users said that the control interface often left them confused about how they got there:

> u3: With the web page [control] method, I quickly got off topic and had to keep multiple tabs open.
>
> u6: [MICE] better shows how papers are related and shows how I got to them.
>
> u2: [MICE] allowed me to traverse through documents while seeing where I was in relation to my past clicks. Whereas the [control] method required me to click the 'back' button anytime I wanted to backtrack on links.
>
> u4: Seeing how papers reference each other was much simpler in the tree view, as opposed to relying on memory and wondering how I got to the current paper from where I started.

3) *Supports comparison.* MICE supported knowledge formation that users thought would be missed while using the control interface:

u7: It is easier to see all the surrounding papers, the ones cited by the paper, referenced by the paper, and the surrounding citations.

u3: MICE definitely gave much more useful information than did the web page [control] method. Each factoid linked directly to other papers that shared some similarity through that particular fact.

4) *Integrated view mitigating disorientation.* Students found MICE's integrated view to be valuable and useful. Student u7 found MICE's visualization of cycles helped him understand which papers he had already seen.

u10 : With MICE, I was able to see more diverse papers in the same viewing space, ... I discovered even more interesting papers from other topics. With the [control] method, interesting papers were more narrow in topic. I had to navigate further to find the next set of interesting papers.

u1: MICE seemed quicker. I like using tabs for doing broad searches like this, but being able to see all the relations on one screen is very useful.

u7: The red line linking the same paper... [helps you] see what papers you have already looked at.

Qualitative data further confirm that metadata works as a form of document summary. The concise representation of metadata summaries helped students read large amount of information and rely less on visual memory. Relationships between previously and newly browsed papers were visible through metadata fields, helping students keep their browsing sessions on track.

Overall, MICE helped students understand context, browse related papers, and build knowledge through citation chaining. Participants preferred MICE for the exploratory browsing task. The time they spent in MICE instead of in the control interface shows that the metadata effectively summarizes digital library entries. The results show that the present interface supports exploratory browsing, while maintaining context for the user.

Question	Rating μ	p
(interesting) Which method helps you better find interesting papers along the citation chain?	1.46	.009
(overview) Which method helps give you a better sense of the referred or cited papers, before you actually read the paper?	1.70	.002
(overall) Which method do you prefer to use, overall?	1.46	.007
(citations) Which method is easier to use for citation chaining?	2.70	$< .001$

Table 2: Mean user ratings on a scale from -4 (strongly preferring control interface) to 4 (strongly preferring MICE), and t-test statistics.

4.4 Implications for Design

We need to discover new methods for making the world's vast, growing information resources more valuable to humanity. Consistently structured metadata representations of widely-used web documents enable summarization, usable presentation, and exploratory browsing experiences that maintain context. From our experiences with the metadata type system and the MICE interface, we derive implications for designing and engineering exploratory browsing and search interfaces supporting open-ended tasks:

Use metadata to represent summaries of web documents. Metadata summaries provide unique value to users browsing large collections of documents. Fields can function as facets, facilitating tasks that involve quick scanning, filtering, and comparison of multiple items, such as sorting products by price or finding most cited papers. Hierarchically nested structures enable collapse and expansion of details, helping users to better allocate their limited attention and make sense of information at different levels of abstraction. Representing linked metadata is an expandable field, in which the field name represents the relationship, helps users form mental models of

citation chains, and so to acquire new knowledge.

In the study, participants with MICE spent 2 orders of magnitude less time (.84 vs. 16.43 min on average) viewing digital library pages, showing that metadata provided by MICE effectively summarized the source documents for the exploratory browsing task. While MICE presents metadata in a table-like structure, other interfaces can use different layouts driven by the same underlying metadata types.

Present metadata summaries clearly, concisely, and consistently. Usable presentation of metadata summaries requires clearly presenting abundant details on the wild web, while managing redundancy and noise. Real world metadata is full of details and cross-references. Abundant details reveal the inherent complexity of the world. Details support various contextualized and personalized user tasks. Tufte wrote: "Detail cumulates into larger coherent structures ... To clarify, add detail." [Tufte, 1990] Abundant details, when properly arranged, make full use of human capabilities of processing information, reduce the need of visual memory for switching contexts [Tufte, 1990], and thus are essential to usable presentation of metadata summaries. In the MICE study, details such as references and citations, which are often absent in prior approaches to metadata summaries [Dontcheva et al., 2006] [Center for History and New Media at George Mason University, 2006], enable citation chaining.

On the other hand, inevitable redundancy and noise in real world metadata distract users' inherently limited cognition. The more information that is presented, the more cognition is consumed [Simon, 1971]. Presentation hints can focus metadata displays toward usability. For example, DOIs [International DOI Foundation, 2018] and URLs [Berners-Lee, 1994] of browsed papers are more useful to the machine than the user. Presentation hints thus guide the dynamic exploratory browsing interface to avoid direct display of this information to the user, while using it as the destination of a hyperlink on the title field.

MICE provides a clear and concise presentation of bibliographic information, removing visual noises and clutters in the original webpage. Participants found this useful. Moreover, MICE provides a consistent presentation across supported websites. MICE presents papers from ACM,

IEEE, and ScienceDirect in a similiar way. This can further reduce the cognitive burden for users to recognize useful information on webpages from multiple sources.

Operate on popular websites and heterogeneous information types, to support real scenarios. Popular websites are, inherently, repositories of information that matters to people. The Semantic Web approach assumes that they will be published using standards for machine-understandable, linked data. Based on this assumption, many Semantic Web applications treat metadata as the result of preprocessing performed in advance. SPARQL queries then retrieve metadata for presentation. Alas, many useful web sites publish only semi-structured HTML, with human-oriented markup and styles, rather than RDF, OWL, or microdata. While a WWW 2007 paper articulated the need to connect semantic web and Web 2.0 approaches [Ankolekar et al., 2007], `programmableweb.com` shows that six years later, RDF plays a role in less than 1% of registered APIs.

The present research enables extraction and presentation of metadata from a wide range of popular web sites people use to find citations, including ACM Digital Library, IEEE Xplore, and CiteSeerX. The MICE interface addresses multiple types of information, such as papers and authors. This is important for building interfaces that provide immediate value to users: we need to support *real world* scenarios, which can involve multiple sources and heterogeneous information types. Working with popular and useful web sites also enables investigation of real world use cases, which leads to holistic, deep understanding of people's practices with web information. This is crucial for driving research into the design and engineering of interactive information systems. Next chapter will present details on how we support dynamic extraction and presentation of metadata from popular web sites.

5. BIGSEMANTICS: A LANGUAGE, TYPE SYSTEM, AND ARCHITECTURE FOR WEB SEMANTICS *

Our study with established investigators shows that, for scholarly research ideation, researchers engage in what we identity as *curating* prior work. Through curation, researchers make sense of and reflect on elements and their relationships, which helps develop new ideas. We make the following observations, which become the motivation for designing and developing BigSemantics, the underlying technology that enables MICE:

1. When thinking about abundant, heterogeneous information, which represents things that become significant in their experiences, people naturally categorize things, based on common attributes [Bowker and Star, 2000]. For example, researchers categorize scholarly publications into different types, such as conference papers, journal articles, and books. This categorization helps participants think of different information, e.g., people usually think that journal articles are more systematic, whereas conference papers are more up-to-date. Categorization is a basis for developing more complicated concepts [Lakoff and Johnson, 1999].

   In programming languages, developers use *types* to name, abstract, organize, and communicate categories and concepts involved in the problem they intend to solve [Mitchell et al., 2003; Pierce and Benjamin, 2002]. A type in a programming language specifies how data is structured and can be used. Naming the specification with a type name supports referring to it, such as in functions operating on such data. *Abstract data types*—types associated with a set of intended operations—support information hiding, which hides implementation details and enforces intended uses of data, proves useful for building complex software. In these

*Edited reprint with permission from "Metadata Type System: Integrate Presentation, Data Models and Extraction to Enable Exploratory Browsing Interfaces" by Yin Qu, Andruid Kerne, Nic Lupfer, Rhema Linder, and Ajit Jain, 2014. In *Proceedings of the SIGCHI Symposium on Engineering Interactive Computing Systems*, pages 107-116, DOI: `https://doi.org/10.1145/2607023.2607030`. Copyright 2014 by Qu, Kerne, Lupfer, Linder, and Jain. Publication rights licensed to ACM.

ways, types help developers think about and correctly use data in their programs, and can function as documentation that is always up-to-date.

Borrowing the idea of types from programming languages research, in the context of using metadata in applications and interfaces to support meaningful user tasks, we define *metadata types* as specifications of metadata, on how it is structured, how it can be obtained, and how it should be presented to users. With the notion of metadata types, we hope to help developers think about what semantics are involved in their applications, how they are operated on, and how they can be reused, just like thinking about abstract data types when they program. We expect that providing metadata types as a tool for thinking about real world information in more abstract ways will streamline the process of developing applications presenting semantics.

2. Researchers work with abundant information, with rich details, from many sources on the web, rather than limiting themselves to certain types and sources. For example, people get bibliographic information from search engines, digital libraries, authors' personal websites, reference lists in known papers, discussions with colleagues, etc. Interfaces supporting researchers need to support popular types and sources. This presents at least two challenges:

   (a) Most sources researchers use are designed for human consumption, and do not publish structured, machine readable semantics in formats such as RDF;

   (b) Websites can drastically change their design. New sources and information types may emerge. Interfaces addressing scholarly research are expected to continuously support existing sources and take new, useful sources and types into consideration.

3. Fundamental limitations of human cognition, such as working memory capacity [Cowan, 2001], present further challenge when researchers need to work with abundant information. Many study participants reported difficulties of keeping collected information organized, finding specific pieces of information when they need it, or putting multiple elements together to form a coherent idea. This problem is aggravated by many information sources

using different designs and being visually cluttered with advertisements and other UI elements. Interfaces addressing scholarly research need to help users allocate their limited cognitive resources, focus on information pertinent to their ideation tasks at hand, and reduce extraneous cognitive load for identifying useful information from varying designs and visual clutters.

Our study with MICE users shows that, metadata—which we defined in Chapter 4 as contextual information describing real world entities in detail—can be used as valuable summaries. It is thus reasonable to expect metadata to be useful in scholarly research ideation activities, such as search, browsing, and curation. In search and browsing, metadata provides concise, consistent representations of browsed information, reducing visual clutters from the original webpages, and helping people focus. In curation, metadata provides contextual information and details, supporting discovery of unexpected relationships and development of new connections.

Therefore, to support scholarly research ideation in real world use cases, dynamic exploratory browsing interfaces need to present abundant, rich, heterogeneous, yet concise and consistent web semantics, from multiple sources, to make best and full use of human cognitive capabilities. We need tools to programmatically obtain metadata from popular websites, remove clutters, and present metadata of the same type in consistent and concise forms.

To address this need, we develop *BigSemantics*, a language, type system, library, and architecture for dynamic extraction and presentation of web semantics. Figure 5 shows a procedural overview of how BigSemantics supports dynamic exploratory browsing interfaces, addressing the entire life cycle of specifying / obtaining / presenting metadata:

1. Interface developers author code snippets called *wrappers*, in the *meta-metadata* language provided by BigSemantics, to specify details of *metadata types* valuable to users. A metadata type describes multiple aspects of the semantic information it represents, including its data model, how it can be obtained, and how it should be presented (see Section 5.1 for a definition).

2. Metadata types can be reused, through *inheritance*, forming a type system. BigSemantics comes with a repository of commonly used types, addressing popular websites. New wrap-

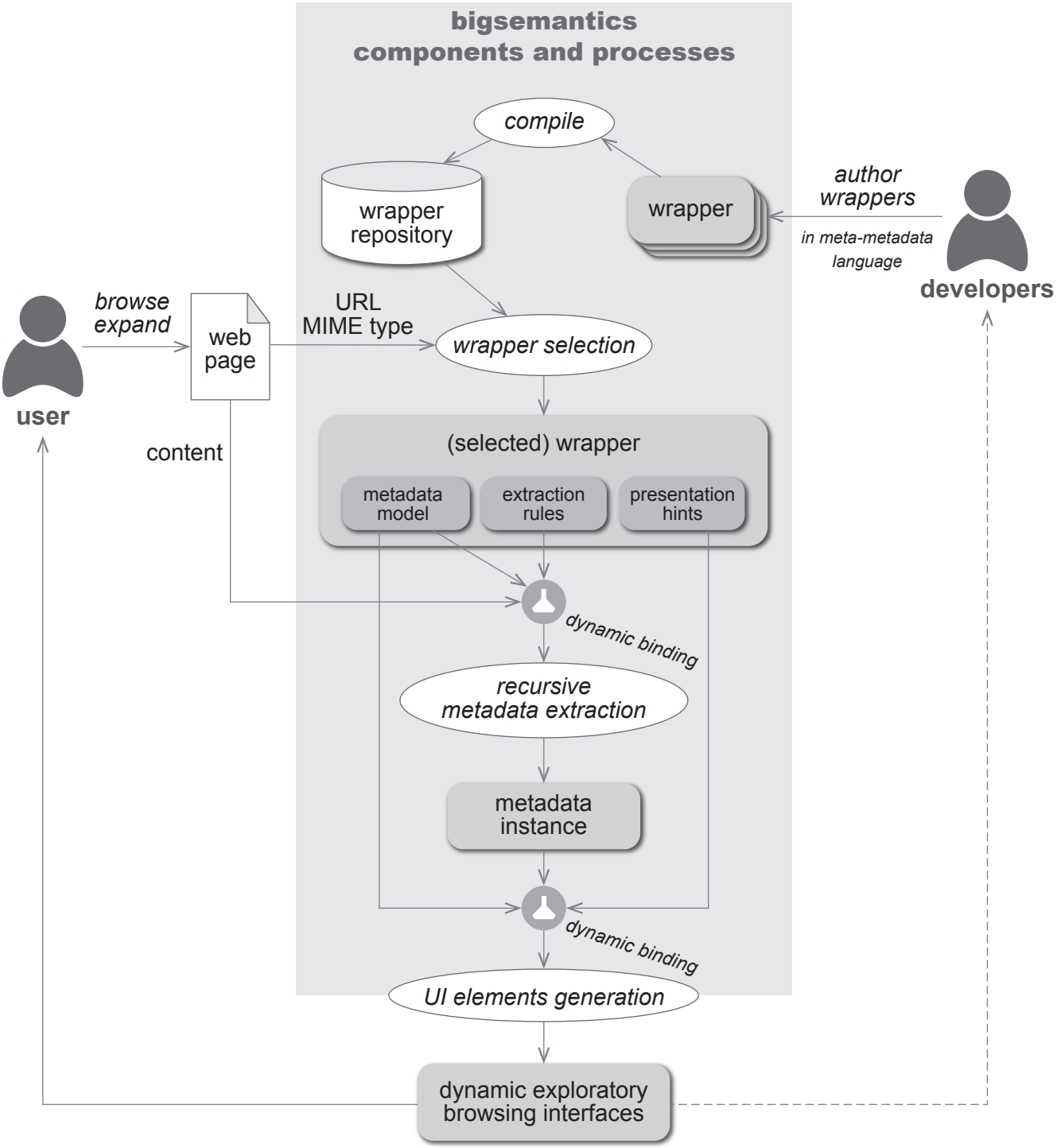

Figure 5: BigSemantics for Dynamic Exploratory Browsing Interfaces: A Procedural Overview

pers can be authored and added to the repository, to support new metadata types and sources.

3. The repository of wrappers is compiled before loaded by the BigSemantics runtime library, to minimize processing time when the application starts. The compilation process takes wrappers in the repository, resolves inheritance relationships, pre-populates fields with inherited nested fields, extraction rules, and presentation hints, generates ad hoc types for particular use cases, resolves generic types, and detects errors.

4. At runtime, when the user encounters a web page, the BigSemantics library uses the URL of the page, and potentially its MIME type and suffix, to select a wrapper from the repository that best suits the incoming page. The BigSemantics library then uses extraction rules specified in the selected wrapper to recursively form a metadata instance out of the incoming web page.

5. Metadata instances formed by the BigSemantics runtime, dynamically bound with the corresponding metadata model and presentation hints, is subsequently used to generate UI elements, resulting in dynamic, concise, consistent, and usable presentations of web semantics.

In this chapter, we first introduce the meta-metadata language, and details on how to author wrappers. We will also take a look at the repository coming with BigSemantics. Then, we explain how the BigSemantics runtime supports dynamic acquisition and presentation of web semantics. The notion of polymorphism plays an important role in this process. We then take a closer look at the use of BigSemantics in supporting an example dynamic exploratory browsing interfaces, MICE. We end this chapter by discussing lessons and implications learned from developing and using BigSemantics.

5.1 The Meta-Metadata Language

With BigSemantics, developers author *wrappers*, which are code snippets written in an XML-based language, *meta-metadata* [Kerne et al., 2010], to specify metadata types. A *metadata type* involves:

```
<meta_metadata name="acm_article" extends="scholarly_article" parser="xpath">
  <selector url_stripped="http://dl.acm.org/citation.cfm" />
  <example_url url="http://dl.acm.org/citation.cfm?id=642611.642681" />
  <filter_location>
    <set_param name="preflayout" value="flat" />
    <strip_param name="CFID" />
    <strip_param name="CFTOKEN" />
    <replace pattern="id=\d+\.(\d+)" to="id=\$1" />
    <alternative_host>portal.acm.org</alternative_host>
  </filter_location>

  <scalar name="title">
    <xpath>//div[@class='large-text']/h1</xpath>
  </scalar>
  <scalar name="location" bibtex_type="URL"/>
  <scalar name="description" label="abstract">
    <xpath>//h1/a[contains(text(), 'ABSTRACT')]/ancestor::h1/...</xpath>
  </scalar>

  <collection name="authors" child_type="acm_author" show_in_snippet="true">
    <xpath>//div[@id='divmain']//a[@title='Author Profile Page']</xpath>
    <scalar name="title">    <xpath>.</xpath>       </scalar>
    <scalar name="location"> <xpath>./@href</xpath> </scalar>
    <collection name="affiliations" child_type="acm_institution_profile">
      <xpath>../..//a[@title='Institutional Profile Page']</xpath>
    </collection>
  </collection>

  <composite name="journal" type="rich_document" show_in_snippet="true">
    <xpath>(//td[contains(text(), 'Published in:')]/../../tr)[4]/td</xpath>
    <scalar name="title">    <xpath>.</xpath>       </scalar>
    <scalar name="location"> <xpath>./@href</xpath> </scalar>
  </composite>

  <collection name="references" child_type="acm_article" show_in_snippet="true">
    <xpath>//h1/a/span[contains(text(), 'REFERENCES')]/ancestor::h1/...</xpath>
    <scalar name="title" layer="20"> <xpath>.</xpath>       </scalar>
    <scalar name="location">         <xpath>./@href</xpath> </scalar>
  </collection>
  <collection name="citations" child_type="acm_article" show_in_snippet="true">
    <xpath>//h1/a/span[contains(text(), 'CITED BY')]/ancestor::h1/...</xpath>
    <scalar name="title" layer="20"> <xpath>.</xpath>       </scalar>
    <scalar name="location">         <xpath>./@href</xpath> </scalar>
  </collection>

  <!-- More fields omitted -->
</meta_metadata>
```

Listing 1: An example wrapper defining a metadata type for ACM Digital Library articles. It reuses an existing type, `scholarly_article`.

- a *metadata model*, which defines the structure of information of this type,
- *extraction rules*, which indicate how information can be acquired from sources, and
- *presentation hints*, which guide how information should be shown to the user.

Grammatically, a wrapper is a `meta_metadata` element with a list of nested fields (Listing 1). The name attribute of every `meta_metadata` element must be unique. There are 3 kinds of fields:

- A *scalar field* defines a typed slot for scalars – values conveniently represented as a string. How the string value should be interpreted is specified by the `scalar_type` attribute. For example, field `year` in wrapper `creative_work` specifies a slot for an integer.
- A *composite field*, such as `rich_media` in wrapper `creative_work`, defines a slot for an instance of a specified `type`. It can also uses the `extends` attribute to derive a new type on top of a base type.
- A *collection field* defines a slot for a set of instances of a common type specified by `child_type`. Similarly, a collection field can use `child_extends` to derive a new type on top of a base type.

Composite and collection fields can further contain nested fields, allowing hierarchical composition of new, complex types.

5.1.1 Extraction Rules

Extraction rules are key to translating regular web pages designed for humans into structured metadata that can be conveniently processed by applications.

To specify extraction rules on a wrapper, the developer first adds the `parser` attribute on a wrapper, to indicate the method used for extraction. Currently, we support 3 methods: XPath for HTML and XML resources, JsonPath for JSON resources, and direct binding for XML resources (when the XML schema aligns with the wrapper's data model). Developers can write their own parsers in JavaScript and add them to BigSemantics.

5.1.1.1 XPath Parser

A majority of wrappers in our current wrapper repository use the XPath parser to extract web semantics from regular, human-oriented HTML webpages. XPath is a standard for specifying the path from the root of the DOM tree to a speicifc node [W3C, 2010]. For robustness, XPath expressions can be specified using semantic features in the DOM, such as element IDs, CSS classes, and semantic tags (e.g. `title`, `article`), which are more likely to remain intact when changes are made to the page.

XPath extraction can be hierarchical. The DOM node corresponding to a composite or collection field can function as the context for evaluating relative XPaths on fields nested in it. In practice, this helps write concise, readable XPaths.

In some cases, the same XPath can be shared in multiple fields. The meta-metadata language allows wrapper authors to define local variables in the scope of the current wrapper to point to DOM nodes, and refer to these nodes using variable names when needed.

Some websites have slightly different templates for the same content. A known example is Amazon, which uses visually similar but structurally different templates for products in different departments. To make authoring extraction rules convenient, the meta-metadata language supports multiple XPath specifications on a single field. The BigSemantics runtime will attempt to evaluate XPaths in the specified order, until a node (for scalar and composite fields) or a non-empty list of nodes (for collection fields) can be located.

5.1.1.2 Field Ops

It is a common need to perform simple operations on extracted content, such as trimming and replacing text. Developers can specify *field ops* on fields to perform these simple operations immediately after web semantics are extracted. Supported field ops are:

- Trimming (a.k.a. stripping) text from one or both ends.
- Prepending / appending text.

- Extracting pattern using a regular expression. Besides taking the part that matches the specified pattern, one can also specify predetermined text or a "capturing group" (a portion from the matched part) as result.
- Replacing pattern using a regular expression. The new text can be either predetermined text or a capturing group.
- Getting a part from the original text. This is similar to pattern extraction, but conveniently supports common use cases, such as "getting the content after foo inclusively".
- Concatenating text from other fields. A circular reference, e.g. field 1 concatenates text from field 2 and 3, but field 2 concatenates text from field 1 and 3, will result in an error.
- Decode URLs. This is useful when the extracted URL was percent-encoded in accordance to RFC 3986, usually because it was used as a query parameter in another URL.
- Getting the content of a query parameter from a URL.
- Setting a query parameter onto a URL.
- Stripping specified query parameters from a URL.
- Stripping all but specified query parameters from a URL.

There can be multiple field ops on one field. In that case, the BigSemantics runtime will perform these ops in the specified order, taking the output from the previous op as the input to the next.

5.1.2 Presentation Hints

Study data shows that presenting concise and consistent metadata is crucial for applications to support users in their exploratory browsing tasks.

The presentation and visualization of web semantics is crucial for interfaces to successfully support user tasks. Thus, it is important to specify aspects of the presentation and visualization

when supporting an information source of web semantics. For example, interfaces supporting scholarly research ideation should present information concisely, to make good use of limited human cognition; clearly, to help the user focus attention on what is expected to be significant; and consistently, to facilitate forming relationships and drawing connections.

The meta-metadata language supports specifying *presentation hints* on fields, to guide how the field should be presented and visualized. For example, the hint `hide` indicates that a field should be hidden from presentation. This is important, because some fields may be useful for the application, but not meaningful to the user, such as internal IDs or sequence numbers. Presenting these fields would be a waste of not only screen estate, but also the user's attention and other cognitive resources.

Another important presentation hint, `navigate_to`, associates a URL field to another text field, creating an anchor. This is needed for the interface to afford hypertextual navigation by a click on the anchored field.

The presentation hint of `layer` uses a number to indicate the importance / priority of fields. At presentation time, fields are ordered by their `layer` attributes. `style` changes the look and feel of a field, for example, if the field should be presented as a title. In the right API environment, such as on the web, the `style` hint can also be used to directly specify CSS.

5.1.3 Inheritance and the Metadata Type System

In objected-oriented programming, inheritance is a mechanism for reusing definitions from an existing structure (e.g. a class) to a new one. In designing BigSemantics, we borrow the concept of inheritance from object-oriented programming, to support reusing metadata types—including their data models, extraction rules, and presentation hints.

When we say metadata type A inherits from metadata type B, we mean that A will reuse the attributes and nested structures already defined for B. In this case, we say that A is a *subtype* of B, or B is a *base type* of A. Through inheritance, a relationship is established between A and B: wherever an instance of type B is expected, an instance of type A can be used. This is called *subtyping polymorphism* [Pierce and Benjamin, 2002], and is crucial for BigSemantics to operate

on and scale up to many metadata types.

A special wrapper attribute, `extends`, is used to indicate inheritance from an existing metadata type when authoring a new wrapper. The value of `extends` should refer to an existing wrapper (the wrapper defining the base type) by name. For example, in Listing 1, wrapper `acm_article` inherits from `scholarly_article` (shown in Listing 2), which in turn inherits from `creative_work` (also shown in Listing 2).

In the subtype wrapper, developers can change data models (e.g., adding new fields), add or change extraction rules, and add or change presentation hints. This works in a "cascading" manner: definitions in the subtype wrapper will be merged onto the inherited attributes and nested structures, and will take precedence. For example, `citations` in `acm_article` inherits from the identically named field in `creative_work`, but changes its type from `creative_work` to `acm_article`; new presentation hint, `show_in_snippet`, is added (Listing 1). This is in many ways similar to how styles are applied and reused in CSS [W3C, 2011]. In CSS, styles for one element are usually applied by a series of rules ordered by the specificity of their selectors: broad selectors with general rules are applied to many elements, to set basic style such as font, whereas specific selectors with particular rules are applied to a small set of elements, to provide control over details. This allows for flexibility while maximizes reuse of style sheets across documents. The styles eventually applied to an element are merged from multiple places. Similarly, in BigSemantics, with the "cascading" manner , one can specify general metadata data models, extraction rules, and presentation hints in the base type, and at specific places where the base type is used, customize it in detail. The eventual metadata type is merged from all the specifications, of various origins.

One common use case for inheritance is to reuse the data models and the presentation hints of one common metadata type, but attach different extraction rules in subtypes to target different information sources. In practice, it is a common pattern for developers to start with a set of base wrappers which define the common data models and presentation hints needed for the application to support user tasks. Then, for each website publishing relevant semantic informa-

```
<meta_metadata name="creative_work" extends="rich_document">
  <collection name="authors" child_type="author" layer="8"/>
  <scalar name="description" label="abstract" layer="9"/>
  <scalar name="year" scalar_type="string" is_facet="true"/>
  <scalar name="overall_rating" scalar_type="string"/>
  <composite name="rating" type="rating" />

  <collection name="references" child_type="rich_document" layer="20"/>
  <collection name="citations"  child_type="creative_work" layer="30"/>
  <collection name="keywords"   child_type="rich_document"/>
  <composite name="rich_media" type="rich_document" >
    <scalar name="title" style_name="normal" />
  </composite>
</meta_metadata>

<meta_metadata name="scholarly_article" extends="article">
  <composite name="source" type="periodical" layer="7"/>
  <collection name="classifications" child_type="rich_document"/>
  <scalar name="pages" scalar_type="string" layer="-1"
          navigates_to="table_of_contents"/>
  <scalar name="citation_count" scalar_type="string" layer="1" is_facet="true"/>
  <composite name="rich_media" label="full text" not_expandable="true"/>
</meta_metadata>
```

Listing 2: Wrapper `creative_work` and `scholarly_article`.

tion, a new wrapper is authored, with specific extraction rules, to extract semantic information from that website. For example, to present bibliographic information, we authored a base wrapper, `scholary_article`, to define the common data model and presentation hints, and created multiple wrappers, including `acm_article` and `ieee_article`, to target specific digital libraries.

Currently, BigSemantics only supports single inheritance, meaning that one wrapper can extend only one other wrapper. One of the considerations for this design was to make it easy to map metadata types defined in wrappers and their inheritance relationships to then popular programming languages, such as Java and C#. Future versions of BigSemantics will explore *mixin*-based approaches [Bracha and Cook, 1990] to multiple inheritance, and potentially focus on JavaScript.

Through inheritance, wrappers developers authored formed a large repository [Interface Ecology Lab, 2016], addressing a wide range of metadata types, websites, and use cases. Figure 6 shows a portion of the repository.

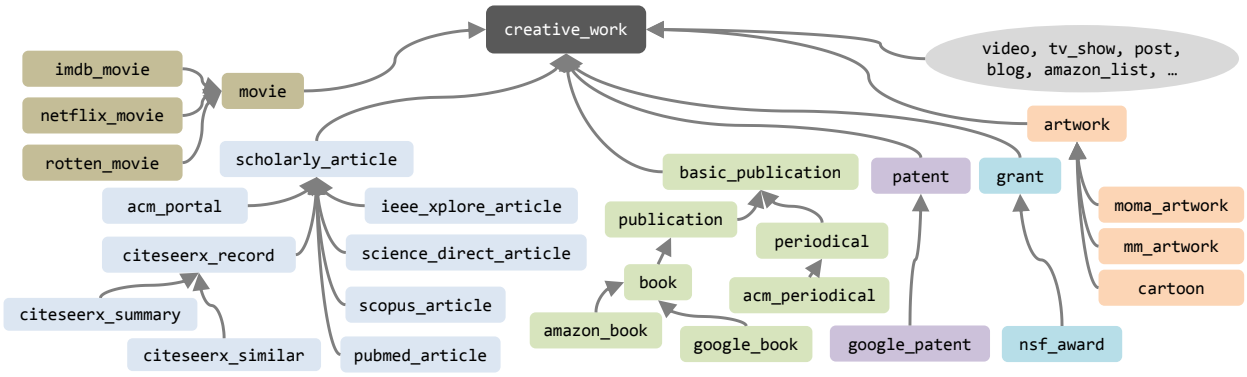

Figure 6: A subset of creative work types supported by the current BigSemantics wrapper repository, and their inheritance relationships. Arrows denote "is-a" relationships.

5.1.3.1 Resolving Inheritance Relationships

To minimize the time needed for processing wrappers and their inheritance relationships, BigSemantics comes with a compiler that collects all the wrappers in the repository, resolves inheritance relationships, pre-populate fields with inherited attributes and nested structures, generates ad hoc types when needed, detects errors, and produces a single file containing the entire repository.

Resolving inheritance between two wrappers is a recursive process. The algorithm recursively processes each field and fields nested inside it. For a field, the algorithm copies definitions from two sources: the wrapper defining the type used for this field, and the *super field*, i.e., the field with the same name and structurally at the same position in the base type. Both sources can specify inheritable attributes and structures. For example, the field `authors` in `acm_article` needs to copy attributes and nested structures from both: 1) the type-defining wrapper, `acm_author`, which specifies what fields should an `author` metadata object has, and 2) the super field, `authors` in `creative_work`, which specifies presentation hints for presenting author lists for creative works.

Listing 3 shows a simplified overview of the inheritance algorithm. The inheritance resolving algorithm take a field and its super field as input. Whole wrappers are treated as composite fields. The algorithm can be split into two major parts, merge and inherit. When the algorithm processes

```
// Merge attributes and nested structures into field from superField.
function merge(field, superField) {
  var names = union(field.nestedFieldNames(), superField.nestedFieldNames())
  // First merge attributes, key to finding types and super fields.
  for (var name in names) {
    var f0 = superField.nestedField(name), f1 = field.nestedField(name)
    if (f0 != null && f1 == null) {
      // f0 is an existing subfield not changed during inheritance.
      field.addNestedSuperField(f0)
    } else if (f0 == null && f1 != null) {
      // f1 is a new subfield.
      f1.setParentField(field)
    } else {
      // f1 is a subfield inheriting from f0, with changes.
      f1.setParentField(field)
      f1.mergeAttributesFrom(f0)
    }
  }
  // Inherit nested structures.
  for (var name in names) {
    var f0 = superField.nestedField(name), f1 = field.nestedField(name)
    f1.setProcessing(true);
    if (f1 != null) inherit(f1, f0);
    f1.setProcessed(true);
  }
}

// Inherit attributes and nested structures into field from its super field.
function inherit(field, superField) {
  // Terminate recursion if reaching scalar fields or fields in processing.
  if (isProcessing(field) || field.isScalar()) return
  // Make sure that the super field is processed. Recurse if needed.
  if (!isProcessed(superField)) inherit(superField, superField.superField())
  field.scope().addSuperScope(superField == null ? null : superField.scope())
  // Find the type-defining wrapper. Create an ad hoc one if needed.
  var typeWrapper = findOrCreateType(field, field.scope())
  // Merge attributes and nested structures.
  merge(field, typeWrapper.asCompositeField())
  if (superField != null && isTypeCompatible(field, superField)) {
    merge(field, superField)
  }
}
```

Listing 3: A simplified overview of the inheritance algorithm BigSemantics uses to pre-process the wrapper repository, written in JavaScript.

nested structures, the two parts call each other, forming a recursion. Descriptions of the two parts:

- Function merge, which copies and merges subfields from the super field into the field. This function goes through the list of subfields twice. The first pass only copies and merges attributes. The second pass copies and merges nested structures. Separating the two passes prevents infinite loops, as the algorithm checks for copied or merged attributes to determine if a field is in processing.

  For each subfield, function merge takes one of the following actions:

  - If the subfield exists in the super field and is not changed in the current field, the algorithm puts a reference of it into the current field.
  - If the subfield is newly defined, the algorithm calls inherit without setting the second parameter (the super field).
  - If the subfield exists in the super field and is changing in the current field, the algorithm merges attributes from the one in the super field into the one in the current field, and calls inherit to recursively process it. This produces a final, changed version of the subfield inside the current field.

- Function inherit, which prepares the input field and super field before copying and merging definitions. Specifically, this function:

  - makes sure that the super field is processed for inheritance;
  - finds the metadata type applicable to the values of the field (e.g., for field `authors`, the applicable metadata type is `author`);
  - allows the field to access names in the super field's scope (this is necessary because the super field can use an ad hoc metadata type);
  - creates ad hoc metadata types when specified by the developer.

After preparations, inherit calls merge twice, to copy and merge definitions from the type defining wrapper and the super field—we have explained that both sources can specify inheritable attributes and structures.

The two functions call each other, forming a recursion. This recursive inheritance process enables flexible, "cascading" customization of data models, extraction rules, and presentation hints.

To see how the algorithm works, let's look at an example: in Listing 1, wrapper `acm_article` inherits from wrapper `scholarly_article`, which further inherits from wrapper `creative_work` (see the type hierarchy in Figure 6). The field `citations` in `acm_article` inherits from `citations` in `creative_work`. That is, the former reuses attributes and nested structures already defined for the latter. To distinguish the two, we call the existing field (`citations` in `creative_work`) the *super field*, and the new field (`citations` in `acm_article`) the *subfield*, or simply the *field* when it is unambiguous in context.

The super field, `citations` in `creative_work`, is a collection field whose values are of the type `creative_work`. It is defined as:

```
<!-- Inside wrapper creative_work -->
<collection name="citations" child_type="creative\_work" layer="30"/>
```

The subfield, `citations` in `acm_article`, while inheriting from the super field, makes modifications, such as changing `child_type` from `creative_work` to (a more specific type) `acm_article`, adding a presentation hint `show_in_snippet`, and attaching new XPaths to itself and nested fields (e.g. `title` and `location`) for extraction:

```
<!-- Inside wrapper acm_article -->
<collection name="citations" child_type="acm_article" show_in_snippet="true">
  <xpath>//h1/a/span[contains(text(), 'CITED BY')]/ancestor::h1/...</xpath>
  <scalar name="title" layer="20"> <xpath>.</xpath>       </scalar>
  <scalar name="location">         <xpath>./@href</xpath> </scalar>
</collection>
```

The inheritance resolving algorithm will first merge attributes from the super field into the current field. The result will be equivalent to:

```
<!-- Inside wrapper acm_article -->
<collection name="citations" child_type="acm_article" layer="30"
            show_in_snippet="true">
  <xpath>//h1/a/span[contains(text(), 'CITED BY')]/ancestor::h1/...</xpath>
  <scalar name="title" layer="20"> <xpath>.</xpath>       </scalar>
  <scalar name="location">         <xpath>./@href</xpath> </scalar>
</collection>
```

Attribute `layer` is copied from the super field, whereas attribute `show_in_snippet` is kept unchanged. Both sources specify the attribute `child_type`, whose values are in practice checked to ensure type compatibility: the type for the current field (`acm_article`) must be the same as or a subtype of the type for the super field (`creative_work`).

After resolving type-related and other attributes, the inheritance resolving algorithm then finds the wrapper defining the type for the field's values (`acm_article` in this example), and copies field definitions from the type-defining wrapper into the field. The result will be equivalent to:

```
<!-- Inside wrapper acm_article -->
<collection name="citations" child_type="acm_article" layer="30"
            show_in_snippet="true">
  <xpath>//h1/a/span[contains(text(), 'CITED BY')]/ancestor::h1/...</xpath>
  <scalar name="title" layer="20"> <xpath>.</xpath>       </scalar>
  <scalar name="location">         <xpath>./@href</xpath> </scalar>

  <!-- Below are structures copied from wrapper 'acm_article' -->

  <scalar name="description" label="abstract"></scalar>

  <collection name="authors" child_type="acm_author" show_in_snippet="true">
    <scalar name="title"></scalar>
    <scalar name="location"></scalar>
    <collection name="affiliations" child_type="acm_institution_profile">
    </collection>
  </collection>

  <composite name="journal" type="rich_document" show_in_snippet="true">
    <scalar name="title"></scalar>
    <scalar name="location"></scalar>
  </composite>

  <collection name="references" child_type="acm_article" show_in_snippet="true">
    <scalar name="title" layer="20"></scalar>
    <scalar name="location"> <xpath></scalar>
    <!-- More fields omitted -->
  </collection>

  <!-- Field 'citations' is a recursive reference to the same field -->
  <collection name="citations" child_type="acm_article" show_in_snippet="true">
```

```
  </collection>

  <!-- More fields omitted -->
</collection>
```

Notice that the field contains a reference to itself (the nested `citations`). This is because the data model itself is recursive: a citation can have its own citations. The inheritance resolving algorithm further processes each nested field, recursively. The check on if a field is processed or undergoing processing (with helper functions isProcessed and isProcessing in the code listing) prevents infinite loops in this recursion.

This "cascading" process can be arbitrarily deep, i.e., one can specify attributes, extraction rules, and/or presentation hints on a deeply nested field. This enables flexible customization while supporting reuse of data models, extraction rules, and presentation hints. Inheritance in a typical object oriented programming language, such as Java, only allows for specifying details for immediately fields; if you want to change the type of a deeply nested field, or attach new annotations to it, you will need to create a series of new types to do so. For example, if you want to change the extraction rule on the `title` field for each one of `authors`, you just need to write out how the field is nested (e.g., writing out `authors` and then `title` inside it), and specify the new rule on it.

5.2 The BigSemantics Runtime

The BigSemantics runtime loads the (compiled) wrapper repository, handles extraction of web semantics from incoming webpages, and produces metadata instances as native objects for presentation. When processing an incoming webpage, BigSemantics will match its URL against all the *selectors* from wrappers in the repository, picking the wrapper whose selector fits best. The selector specification includes a URL pattern, in the form of hosts, paths, or regular expressions, and an optional list of required / forbidden query parameters and values. For example, the selector in Listing 1 specifies a URL path (the part without query parameters and fragments) for ACM Digital Library articles. After picking a best-matching wrapper, BigSemantics uses rules specified in the `filter_location` element (Listing 1) to transform and normalize the location of the incoming

webpage, for cross referencing. Common transformations include stripping certain query parameters (e.g. session id parameters, like `CFID` and `CFTOKEN`) and replace alternative hosts with the canonical one.

The BigSemantics runtime then recursively extracts web semantics using the specified parser, such as the XPath parser, resulting in metadata instances, which are native objects that can be further processed and presented by the application. To further facilitate presenting semantic information in useful forms to the users, the BigSemantics runtime can also associate extracted metadata instances with the corresponding wrapper—carrying valuable presentation hints—using the algorithm in Listing 4.

The key feature that enables BigSemantics to support dynamic interfaces is *document subtype polymorphism*. In programming languages, *subtype polymorphism* allows for general functions to

```
function pairFieldsWithMetadata(fields, metadata) {
  var pairedFields = []
  var fieldValue = metadata[field.name]
  for field in fields {
    switch (field.type) {
      case 'scalar':
        pairedFields.push(field, fieldValue)
        break
      case 'composite':
        pairedInternals = pairFieldsWithMetadata(field.nestedFields,
                                                 fieldValue)
        pairedFields.push(field, pairedInternals)
        break
      case 'collection':
        for (var i = 0; i < fieldValue.size; i++) {
          pairedInternals = pairFieldsWithMetadata(
            field.nestedFfields,
            fieldValue[i])
          pairedFields.push(field, pairedInternals)
        }
        break
    }
  }
  return pairedFields
}
```

Listing 4: The recursive algorithm MICE uses to pair field metadata instance data structures with wrapper fields.

operate on instances of different subtypes of a common (base) type, enabling different behaviors at runtime and promoting reuse. In BigSemantics, *document subtype polymorphism* is a key to addressing heterogeneous information types and sources. The runtime provides general functions, such as metadata extraction and presentation, which operate on the general base type `document`. The type system and runtime then *polymorphically* operate on subtypes of `document`, such as `scholarly_article` and `amazon_product`, to extract heterogeneous metadata with different structures and contents, and consistently display them with rich presentation. This polymorphism is operationalized by dynamic bindings of documents to types integrating data models, extraction rules, and presentation hints, and the invocation of extraction and presentation functions (Figure 5). As new metadata types are introduced (by authoring new wrappers), dynamic exploratory browsing interfaces building upon the type system, such as MICE, immediately support them.

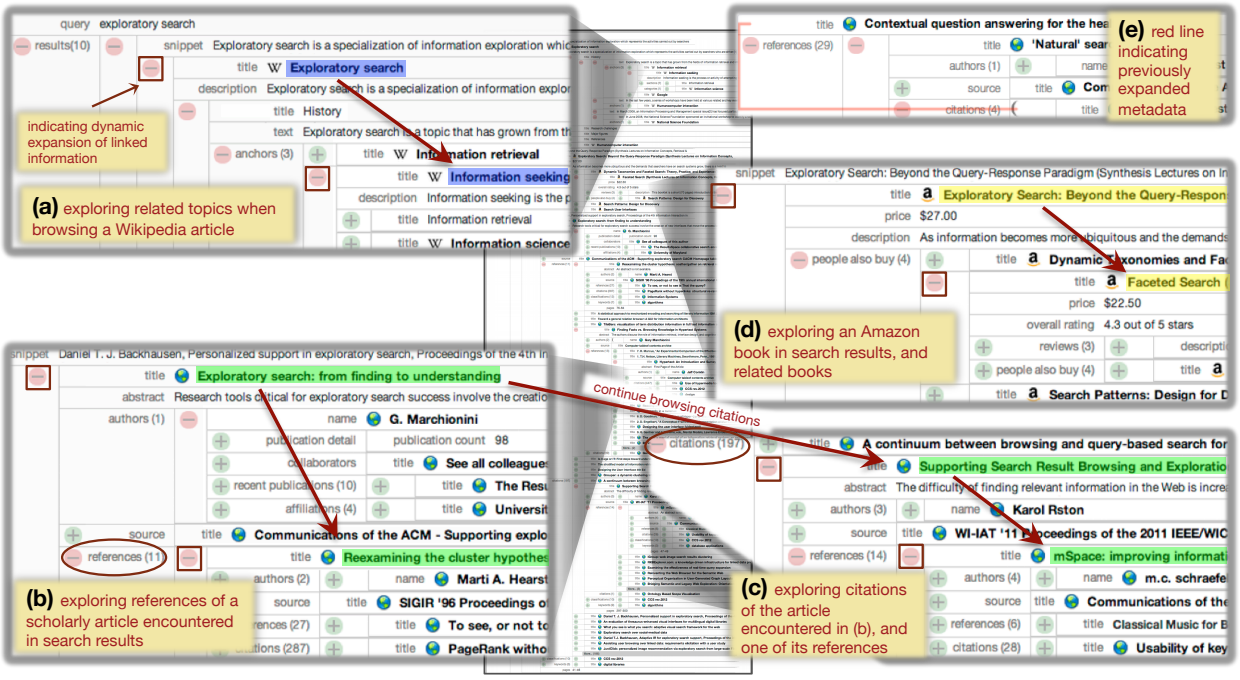

Figure 7: Copy of the overview of Yin's exploratory browsing with MICE, for recap. Snippets show close-up views of her session. Arrows denote browsing linked information.

5.3 Example: Using BigSemantics in MICE

We use MICE as an example to show how BigSemantics can support modeling, extraction, and presentation of metadata in web applications. As a recap, we copy the illustration of Yin's exploration session here in Figure 7.

5.3.1 Representing Documents as Metadata

Limits in human cognition form the basis of a need to concisely and consistently represent documents. In Yin's exploratory browsing session, he encounters diverse documents, such as articles, author profiles, and books. Some documents contain nested structures, such as a long list of citations. Presenting original documents with all the information in one context could overwhelm, since working memory is limited [Cowan, 2001]. To mitigate this, we use BigSemantics to summarize documents as *metadata*. Nested structures, such as citation lists, are broken down into constituent sub-objects, which the user can collapse and expand to focus use of attention and display. Figure 7b shows metadata for a scholarly article, e.g., title, authors, and references.

We use metadata types as a conceptual instrument to address categories people naturally develop when they work with abundant, rich information. The BigSemantics metadata type system specifies types in code blocks called *wrappers*. Figure 8 shows example wrappers used in Yin's scenario. Wrapper `creative_work` specifies a common type for creative work, which includes *fields* such as `year`, `references`, and `rating`. The type system supports three kinds of fields: scalar, composite, and collection. In wrapper `creative_work`, field `references` specifies a reference list in which each reference must be an instance of `document` (or its subtypes by *polymorphism*, which we will explain later), and field `citations` specifies a citation list in which each citation is an instance of `creative_work`. Composite and collection fields can represent relationships between linked metadata, as `references` and `citations` do.

The type system supports *inheritance*, denoted by attribute `extends`, for reusing and extending types. For instance, as a form of creative work, we derive a type, `scholarly_article`, that inherits from `creative_work`, adding new fields such as `source` and `keywords` (Figure

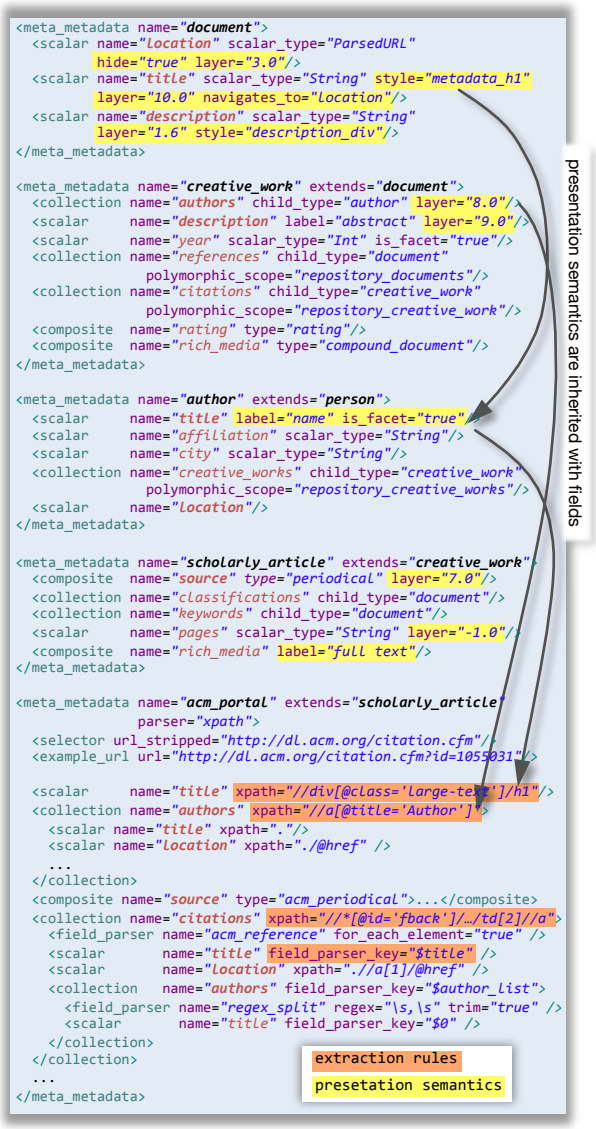

Figure 8: Type inheritance and referencing in example meta-metadata wrappers from Yin's scenario.

8). Wrapper `acm_portal` further subtypes `scholarly_article` to extract metadata in the general scholarly article data model from a specific source (the ACM Digital Library). A common practice is to define a data model in a base type and use it for source-specific extraction in subtypes. The type system defines a common base type, `document`, for general web pages, which includes a `title`, a `location` (the URL), and a `description`.

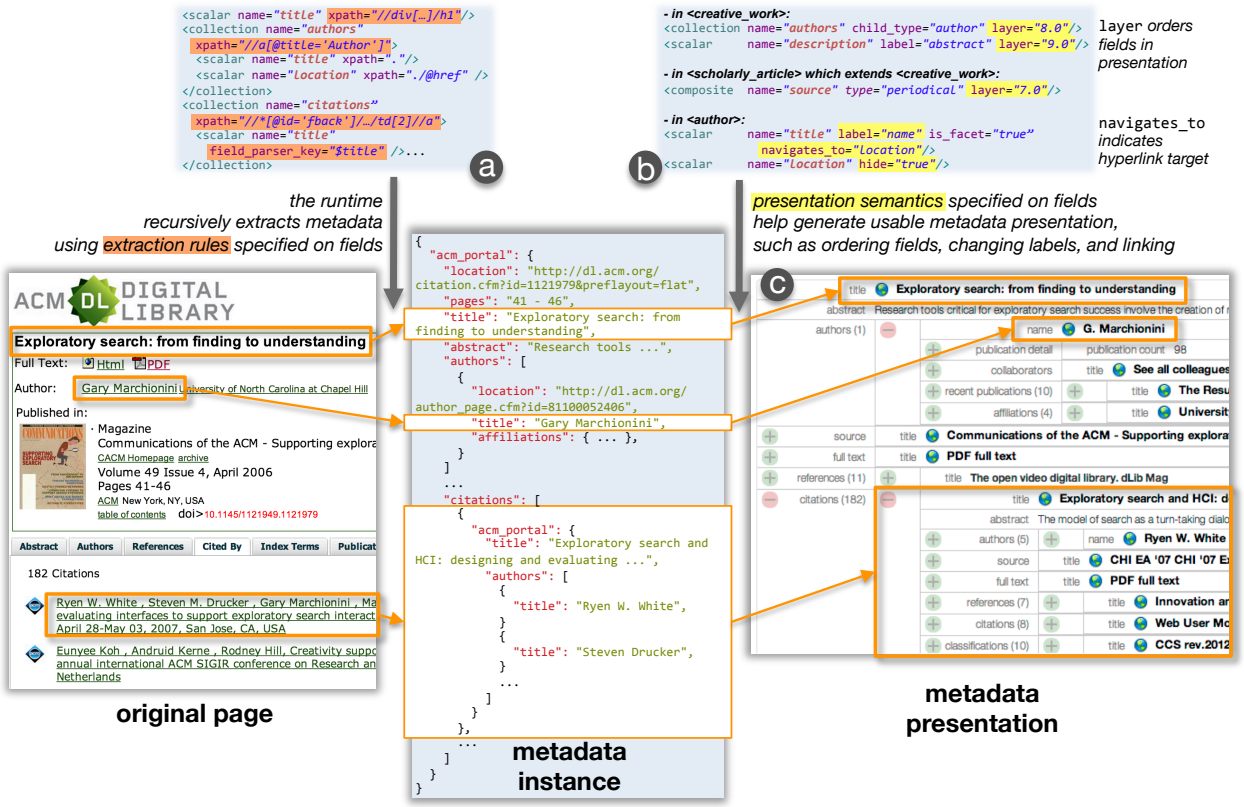

Figure 9: Metadata type system: semantic view. Types drive extraction from the source web page. The type is then joined with the resulting instance (seen as JSON), to drive presentation.

5.3.2 Extracting Heterogeneous Metadata From Documents

A major difficulty with representing documents as metadata is that many popular, useful web sites do not publish metadata. The metadata type system addresses this by actively extracting metadata of heterogeneous types from regular HTML pages published by these sites.

The extraction process begins when the user encounters a document. In Yin's case, when he clicks on the plus button to expand the encountered ACM Digital Library article, MICE makes a request to an underlying typed metadata service to extract metadata for that article. Since Yin could encounter many types of documents, the service needs to select the wrapper appropriate for the requested document. This is enabled by matching URL or mime-type features using *selectors* defined in wrappers. In Figure 8, the wrapper `acm_portal` specifies a selector for ACM Digital

Library articles, with a URL pattern as the feature. Once matched, the wrapper associated with a selector is bound to the document for extraction.

After binding the wrapper `acm_portal` with Yin's encountered article, the runtime uses *extraction rules* integrated with data model fields to extract metadata from the document. Extraction rules can include (1) XPaths which operate on the HTML DOM tree, (2) names that directly map to elements in XML or JSON documents, and (3) regular expressions for pattern matching and filtering. Figure 8 shows example extraction rules (XPaths) for extracting article metadata from ACM Digital Library articles.

Algorithmically, the extraction process first instantiates an empty metadata object of the selected type, then populates the instantiated metadata object with extracted information by iterating over data model fields. For each field, the integrated extraction rules are used to acquire information from the document (Figure 9a). For a scalar field, the extracted representation, a string, is converted into a value of the specified scalar type, such as integer or URL. For a composite or collection field, the process recursively instantiates and populates sub-object(s), using contextual DOM node(s) located by the extraction rule specified in the declaration of the encompassing composite or collection field.

In Figure 9a, the XPath on `citations` matches a list of contextual nodes, each of which corresponds to an anchored, formatted citation string (framed in the figure) in the original page. Formatted citation strings are parsed into key-value pairs, such as authors, title, and publication venue, using a field parser for ACM reference formats. Values are then assigned to the fields of a nested `creative_work` sub-object, such as `title` and `authors`, by `field_parser_key`. The anchor destination of a citation is extracted using a relative XPath and bound to the sub-field `location`, making the citation sub-object a pointer to linked metadata. Recursively extracting sub-objects is key to supporting nested or linked metadata of types, and experiences such as collapsing, and expanding details. The integration of data models and extraction rules enables this practical, field-by-field, recursive algorithm to derive rich metadata for heterogeneous types.

5.3.3 Heterogeneous Metadata and Presentation Hints

In his task, Yin explores Wikipedia articles, research papers, and Amazon books. For Wikipedia articles, he follows links embedded in paragraphs to related concepts. For research papers, he uses references, citations, and authors to chain to related significant research. For Amazon books, he reads reviews to get others' opinions. Exploratory browsing involves encountering metadata of heterogeneous types. Each type may require *rich* presentation tailored to its specific structures and relationships, to make good use of the user's attention.

The BigSemantics metadata type system uses *presentation hints* to address this heterogeneity. Integrated with data model fields, presentation hints specify how a particular field in a particular type should be presented. presentation hints reference CSS classes to situate the details of presentation in abstractions, such as `metadata_h1`, separating low-level details and parameters, such as fonts, where designers can customize them. We developed a set of simple, yet effective presentation hints, including hiding, ordering, positioning, collapsing, expanding, hyperlinking, concatenating, and changing labels for fields. In Figure 9b, `layer` decides the order of fields in presentation, and `navigates_to` specifies hyperlinking the field to a destination indicated by another field.

Presentation hints can be inherited along with data model fields, and overridden as needed, promoting reuse. For example, `layer` specifications in wrapper `scholarly_article` will be inherited by subtypes such as `ieee_xplorer` and `acm_portal`, if not explicitly overridden. Thus, the field order specified in the base type `scholarly_article` will automatically apply to metadata extracted from any of these digital libraries.

Interfaces can render the same presentation hints in different, yet consistent ways, to meet situated needs. The example, MICE, provides a default hierarchical HTML5 rendering, which will be explained in the next section.

5.3.4 Recursive Expansion of Heterogeneous Metadata

Being able to navigate to linked information with one click is crucial for web usability. By providing previously unanticipated information that evolves the user's understanding and information needs, links function as the basis for exploratory browsing, seeking, berrypicking, and thus, scholarly research ideation. Dynamic exploratory browsing interfaces must support such encounters with linked information, while maintaining context.

MICE uses *recursive expansion of heterogeneous metadata* to address this. A link, such as a citation, is initially presented as an abridged metadata object, with only the title; a plus button indicates further information. When the user clicks the plus button, MICE calls the underlying typed metadata service to extract detailed metadata from the linked document. After extraction, the service sends extracted metadata and the corresponding wrapper back to MICE, for presentation. Upon receipt, MICE recursively binds data model fields with extracted metadata values, and then iterates over these fields to generate HTML5 elements for presentation. Interface generation uses presentation hints to customize display for each particular type, including sorting fields, hiding or changing labels, and hyperlinking.

For example, in Figure 9c, the scalar field `title` is presented as a header, anchored to the source ACM Digital Library page, as specified by `navigates_to`. The fields `authors` and `citations` are presented as lists of nested or linked metadata, initially with 10 items and a "show more" button. On expansion, the generated HTML5 elements are injected, replacing the abridged form with a detailed presentation. A sub-object whose `location` field points to a linked document, such as a citation, can be further expanded, which will recursively trigger the information expansion process.

The whole process of selecting the appropriate type, extracting metadata from the document, and presenting metadata with customized visual elements is dynamic, that is, executed in real time as the user encounters the document. Thus, the interface is able to dynamically present heterogeneous information as metadata that can be conveniently collapsed or expanded to the user in real time, while addressing characteristics of particular types.

5.4 Implications

Through our experiences developing BigSemantics and using it in a dynamic exploratory browsing interface, MICE, we derive the following implications for the design of future technologies working with metadata on the web.

Integrate the meta-information of data models, extraction rules, presentation hints, and type selectors to drive effective, usable metadata summary experiences. To provide value to the user, data models alone are not sufficient. Individual metadata summaries must be extracted and usably presented. Metadata data models, extraction, and presentation are inherently intertwined in user experiences. Integrating meta-information of data models, type selection, extraction, and presentation provides a general method for generating rich presentation. The structures of abundant details are expressed through metadata types, providing for consistency and variation, while enabling management of redundancy and noise. Extraction rules recursively acquire pieces of information from the DOM to form typed metadata instances. presentation hints enable hiding, relabeling, reordering, emphasizing, hyperlinking, expanding, collapsing, and concatenating fields, to generate type-specific rich presentations, addressing diverse use cases from the ACM Digital Library to Amazon and beyond.

The integrative metadata type system, with inheritance, helps scale metadata extraction and presentation to many information sources and types. The selector mechanism automatically picks the optimal type for each encountered document. This is essential for expanding serendipitously encountered metadata. Exploratory browsing interfaces, like MICE, operate on the base type of document, while using integrated types to drive presentation. For development and maintenance, the type system supports reuse and overriding of data model fields, extraction rules, and presentation hints through inheritance [Booch et al., 2007].

Provides a repository of popular metadata types and sources. Such a repository being available can not only accelerate application development by reusing existing code, but more importantly, it encourages the adoption of semantic web and similar technologies, including BigSemantics. The semantic web community shares a large collection of schemas, in formats like OWL.

However, these schemas are not easy to find. Even one can find appropriate schemas to use, it is as difficult to find RDF data from popular websites. Newer approaches, such as Microdata, attempts to make schemas easier to find (the schemas can be downloaded from `https://schema.org`), and data easier to use (embedded in regular HTML5, on some major websites). The BigSemantics approach addresses both issues. By providing a centralized repository with the software, developers can get started on building applications fairly quickly. The software works with the repository to actively extract metadata from popular websites, without replying on websites to publish data in special formats. Using open source collaboration processes, such as pull requests, we expect developers to be able to share their work on new and improved wrappers.

Currently, BigSemantics comes with only one set of widgets for building application interfaces, which we used to build MICE. Future work will explore opportunities of providing a different widget library, to further accelerate development of interfaces presenting metadata from the web.

Build upon web native / friendly technologies. BigSemantics was originally developed in Java, and used in desktop applications written in Java and C#. Web applications have to talk to the BigSemantics runtime through a web service. It is later ported to JavaScript / TypeScript, and used in web applications, with the help of a Chrome extension to handle network requests. This not only facilitates the integration of BigSemantics into web applications, but also makes BigSemantics more readily accessible to application developers, because they are already familiar with the technologies. This is a lesson for future web semantics technologies: if you want your technology to be accessible and used by web developers, build upon web native / friendly technologies.

6. DISCUSSION *

Our study with principal investigators addresses how researchers work with prior work while developing ideas for their research. Through qualitative data analysis, we found that:

1. *Finding interesting prior work is important.* Every participant had engaged in this activity. Rather than being direct and well structured, their processes of finding interesting prior work are usually exploratory and serendipitous. Unexpected encounters happen.

2. *The process of curating prior work is essential throughout scholarly research*: for inspiring initial ideas, for developing research methods, for interpreting study results, and for writing reports and papers. Curation is not just passive collection: it involves actively making sense of the curated work in its original context, interpreting relationships, and eventually recontextualizing the curated work for use in one's own research.

3. *Participants' practices of finding and curating interesting prior work are highly situated.* Their ongoing projects, interactions with colleagues, research fields, personal curiosity, among many other factors, drive their needs for and methods of finding and curating prior work. They use search engines, digital libraries, personal websites, and many other tools, to seek interesting work. They go to conferences, talk to friends and colleagues, and review paper submissions—all of which may lead to discovery of new, interesting work. They curate prior work in all forms: printouts, PDF files, BibTeX entries, screenshots, notes, etc.

4. *Participants use their prior work curation to stimulate new ideas*, e.g., when conceptualizing projects and writing papers. Researchers need to contextualize their work in relevant research fields, to derive new contributions.

*Part of this chapter is reprinted with permission from "Metadata Type System: Integrate Presentation, Data Models and Extraction to Enable Exploratory Browsing Interfaces" by Yin Qu, Andruid Kerne, Nic Lupfer, Rhema Linder, and Ajit Jain, 2014. In *Proceedings of the SIGCHI Symposium on Engineering Interactive Computing Systems*, pages 107-116, DOI: `https://doi.org/10.1145/2607023.2607030`. Copyright 2014 by Qu, Kerne, Lupfer, Linder, and Jain. Publication rights licensed to ACM.

5. *Researchers experience breakdowns (i.e., failures) with existing computational systems*. One example is when they used typical web browsers to explore interesting prior work, they often ended up with many open tabs and browser windows. The presentation of these tabs and windows usually does not reflect why and how researchers get to or use these resources. The result is that researchers often felt lost in the interface after intensive exploration. Another example is when researchers would like to curate an interesting work, the media of curated elements, and the media of the curation as a whole, are often limited in revealing significant content, visualizing relationships, and supporting note-taking.

These findings are consistent with and extend beyond previous research on information seeking (e.g. [Belkin et al., 1982; Marshall, 1997, 2008; Whittaker, 2011]) and creative cognition (e.g. [Wilkenfeld and Ward, 2001]). Taking creative ideation as a perspective enables us to identify previously underinvestigated gaps (e.g. the creative nature of scholarly research), and the resulting experience breakdowns (e.g. insufficient support for contextual exploration and flexible curation of interesting prior work). Our integrated approach results in a new, interdisciplinary model of scholarly research ideation. The model is inherently limited by the scope of the qualitative study, such as the fact that all our participants are from HCI-related research fields, Nonetheless, the model provides valuable insights into how participants work, and by drawing on common themes shared by multiple participants, it is reasonable to expect that the model has the potential to provide valuable insights into investigating scholarly research ideation with other researchers, in other fields.

Dynamic exploratory browsing interfaces are conceptualized to address one of the breakdowns, to support exploration of interesting prior work while enabling users to maintain context. The study with MICE, an example dynamic exploratory browsing interfaces, shows that metadata semantics, when properly presented, can function as a useful representation of scholarly works, to help researchers see a multiplicity of information and relationships in one context, supporting exploration and sensemaking processes. MICE's concise and consistent representation of bibliographic information not only saves screen real estate, but also reduces the extraneous cognitive load required

to make sense of web pages (which are usually cluttered with banners, navigational links, and ads), and supports easy comparison between information from different sources. The integrated view showing citation relationships helps users see where they are in the exploration process, thus reduces digression and disorientation.

In the rest of this chapter, we will discuss on the implications for design we derived through this work, how our work connects and compares to prior research on the Semantic Web, and the limitations of our study.

6.1 Implications for Design

Through conducting studies on researchers' practices, developing a model of scholarly research ideation, building a new interface, and evaluating the new interface, we derive not only direct findings presented in previous chapters, but also broader implications for the design of future investigations into scholarly research and interfaces supporting it. Other creative tasks involving building upon existing knowledge and prior work can benefit from these implications, too.

6.1.1 Use Ideation as a Perspective to Motivate Investigation

We identify and focus on ideation tasks, because scholarly research is, by its nature, a creative endeavor; participants work to develop new ideas that become research contributions. The focus on ideation tasks helps us connect previously separate topics and fields (such as information seeking, creativity, and art), and motivates the construction of the new scholarly research ideation model. Having new ideas, including those in the mini-c and little-c categories, is a common goal in many other human activities, such as information seeking, reading, writing, and management. Thus, we recommend using ideation as a perspective to motivate investigations in these fields, as such a perspective can provide a contextual nexus leading to new, insightful, interdisciplinary findings.

Our focus on the creative nature of researchers' tasks urges us to draw from fields directly addressing creativity, including creative cognition and art:

1. The Geneplore model supports understanding people's iterative processes of developing ideas;

2. The mini-c perspective enables investigations of scholarly research ideation practices beginning at a personal impact level, and continuing as researchers grow ideas toward interpersonal, professional, and societal impact;

3. The art concept and practice of curation distinctly characterizes the practice of putting together a situated context, in which one can derive new meanings and develop new ideas.

Inasmuch as human creativity involves both cognitive and expressive aspects, we argue that addressing creativity requires an integration of interdisciplinary perspectives.

The new, interdisciplinary model of scholarly research ideation forms an integrative basis for framing investigations on researchers' practices, and suggests research questions and design opportunities for future work. For example, it could be valuable to study how information seeking and curation integrate with each other, and design interfaces for researchers to seek and curate information in one context, to support this integration.

The new model also provides a basis for investigating ideation tasks in other contexts, as the involved actions are not exclusively limited to scholarly research. Examples of ideation tasks include planning a vacation, writing an essay, inventing a service, and designing a campaign. In these tasks, people often need to seek information, make sense of found information, curate significant and interesting elements, relate the elements, and make interpretations, in order to develop new ideas. Future research can investigate how the new model applies to other ideation tasks.

6.1.2 Use Dynamic Exploratory Browsing Interfaces to Support Scholarly Research Ideation

Our study with MICE shows that dynamic exploratory browsing interfaces (DEBIs), such as MICE, support exploring prior work while maintaining context (e.g. keeping track of citation relationships), an important part of scholarly research ideation. The key for DEBIs to work in this task is to make significant information about scholarly articles (presented as expandable fields) and their relationships visible in one context. Making information visible is a well established design principle [Norman, 2002]. Keeping information in one context reduces the extraneous cognitive load needed for switching, e.g., from one browser tab to another. Future interfaces supporting

scholarly research ideation (and potentially other creative endeavors) should use DEBIs to address users' needs for manageable exploratory search and browsing.

6.1.3 Integrate Information Seeking with Curation to Support Ideation Tasks

Previous information seeking research calls for investigations into broader task contexts in which seeking occurs [Blandford and Attfield, 2010; Järvelin and Ingwersen, 2004]. One such task context is scholarly research ideation, in which researchers seek interesting prior work, read and make sense of found information, in order to develop new ideas that become novel and valuable contributions. Similarly, in many other creative, open-ended tasks, such as writing an essay and planning a weekend, seeking information is important, while ideation is the bottomline goal.

The new, interdisciplinary model of scholarly research ideation shows that information seeking and curation are integrated in ideation tasks. We need to curate in seeking, because we need to see, organize, manipulate, and think across found information, to integrate our embodied mind and the external world, in order to develop ideas. Curation, in turn, provides a situated context, in which newly found information is interpreted, related, and transformed.

Transient in-browser curation is a manifestation of integration of information seeking and curation. With transient in-browser curation, participants keep browser windows and tabs—corresponding to potentially useful webpages found in seeking—open as means of collecting intermediate findings. While today's browsers support other forms of curation, such as browsing history and bookmarks, transient in-browser curation stands out with its informality and flexibility: one simply keeps the corresponding browser tab open. With typical desktop browsers, tabs can be organized in windows by dragging. In comparison, to save a found webpage as a bookmark, one needs to explicitly specify which folder the webpage should go into, imposing premature formalism [Shipman and Marshall, 1999]. Consequentially, organizing bookmarks often requires significant efforts. In our study, many participants choose not to use browser history or bookmarks as a regular method for curating interesting prior work.

Support for transient in-browser curation in today's browsers is limited. A browser tab, as an element of a transient in-browser curation, is less effective than clippings, in terms of representing

significant portions in the corresponding webpage. Transient in-browser curation often features a linear medium of assemblage—often a row of open tabs in the browser interface—which provides limited support for representing complex linkages and relationships in hypermedia [Landow, 2006]. Further, when the browser session crashes or is accidentally closed, the transient in-browser curation, containing potentially useful intermediate findings, may be lost. We suggest that future work should investigate new methods of supporting transient in-browser curation, maintaining their spontaneity, while using flexible media to better enable making complex relationships visible.

Writing is another important activity that can benefit from this integration of information seeking and curation. Writing a research paper or proposal is a creative process that eventually integrates contextually connected elements—prior work, interpretations, data, ideas—into a coherent argument. Thus, writing becomes a nexus that connects different aspects and components of scholarly research ideation, and benefits from an environment that better supports engagement with different scholarly research ideation activities at the same time. Study shows that, during writing a research paper, researchers seek relevant prior work to support their argument and want to revisit their prior work curation for inspirations (see Section 3.2 for examples). New tools supporting scholarly writing should address this need, providing the ability to explore (by searching and browsing) relevant information in the context of the writing task, curate interesting findings (along with interpretations), and revisit prior work curation as a whole (again in the context of the writing task) to help stimulate new ideas.

6.1.4 Use a Metadata Type System to Facilitate Representing Schemas and Presentations of Metadata

Types provide an easy-to-understand method for categorizing and organizing concepts involved in the information people use in their tasks. Categorizing things with common attributes into types, such as “conference papers” and “journal papers”, is an important step towards thinking abstractly. Programming languages have long used types as a tool to help developers think abstractly about complex problems, without being overwhelmed by details. For many typed programming languages, the compiler can use type information to automatically detect potential errors in the code.

Typed programming languages have seen remarkable success.

Most developers are already familiar with the concept of types. Thus, it is a plausible choice to borrow the idea of types from programming languages in designing and developing BigSemantics. Other tools we borrow from programming languages research into BigSemantics include *inheritance* and *subtyping polymorphism* [Mitchell et al., 2003]. Inheritance enables code reuse. With the type repository which comes together with BigSemantics, developers immediately have access to a wide range of useful metadata on the web, and can focus on designing their own interfaces and systems using metadata to support user tasks. With the help of URL selectors, extraction rules, and presentation hints, subtyping polymorphism enables dynamically extracting metadata from webpages and presenting extracted metadata in consistent, concise, and usable forms. New (sub)types can be added to the repository to support situated needs. This support for runtime metadata extraction and presentation is important, because exploration of prior work in scholarly research ideation is by nature dynamic—researchers' needs and interests evolve during the process, as in other exploratory tasks ([Bates, 1989; Belkin et al., 1982]). All these design considerations together, BigSemantics provides a viable and easy to use method for working with rich and vast metadata on the web.

6.1.5 Draw from Art Practice

Art explicitly focuses on creative processes and expressions. The creative nature of scholarly research resembles that of art creation. In the 20th century, conceptual art emerged as a new methodology, extending work with prior media, such as painting and photography. Ideas and processes, themselves, were given the function and status of art [Lippard, 1973]. The constructs of curator as artist and art exhibition as art work exemplify the conceptual art method.

Scholarly research will also potentially benefit from borrowing from artistic methods and practices. Many scientific breakthroughs, such as the theory of relativity [Thorne, 1994], quantum physics [Gamow, 1985], and discovery of the DNA double helix (Figure 10) [Olby, 2013], began with outrageous, artistic imagination, breaking from existing doctrines. Association and analogy, which is important in art, are found to promote discovery and creativity [Finke et al., 1992; Gen-

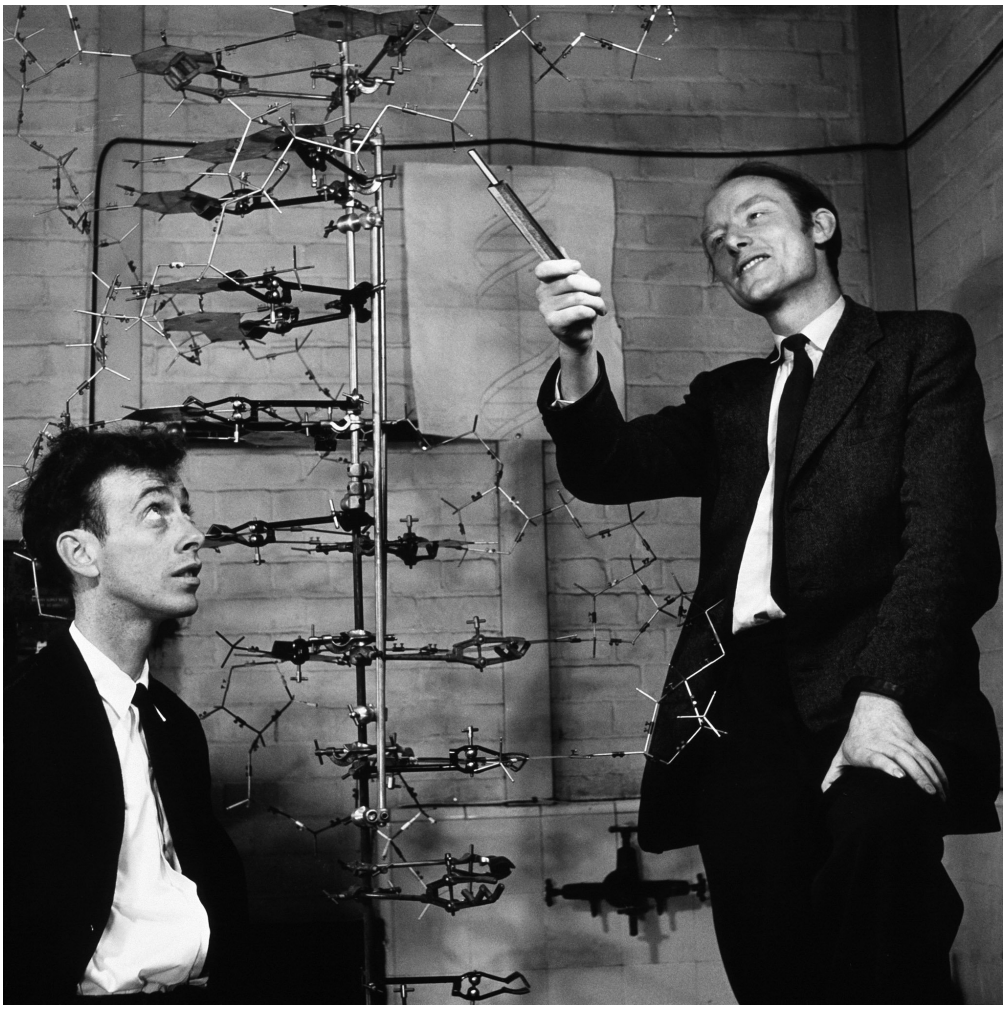

Figure 10: James Watson and Francis Crick with a DNA model at the Cavendish Laboratories in 1953. The double helix structure of DNA was codiscovered by James Watson, Francis Crick, Maurice Wilkins, and Rosalind Franklin. Photo by A. Barrington Brown. Reprinted from [Science History Institute, 2017]

tner, 1983; Nersessian, 1999; Wilkenfeld and Ward, 2001].

Therefore, it is valuable to draw from art concepts and practices to study scholarly research ideation. For example, we use curation, an art-based concept and practice, to characterize scholars' engagement with prior work. This characterization goes beyond individual actions separately addressed in previous information science findings, such as collection and organization, and directly connects to underlying, human-centered desires, goals, and activities—to relate, interpret in

context, recontextualize, and create. Thus, our invocation of curation, as a model component, connects the creative nature of what researchers do, the importance of found objects and assemblage in the arts, and the advent of conceptual art, through which a convergence of art and curation manifested. As such, the present research exemplifies and extends Kerne et al's implication of the value of art, in conjunction with STEM (Science, Technology, Engineering, and Mathematics) [Kerne et al., 2014], to focus and drive innovation. We believe that drawing from art theory and practice will be fruitful for investigations into scholarly research ideation tasks, as well as ideation tasks in other contexts, and furthermore, applications to support the performance of these tasks. New ideas will be stimulated by new mindsets.

6.2 Connecting and Comparing to the Semantic Web

Our approach is closely related to, yet in certain aspects different from the Semantic Web approach [Berners-Lee et al., 2001]. The Semantic Web is a set of standards and technologies that aim to make information machine readable and exchangeable across applications and organizations. A standard at the core of the Semantic Web approach is Resource Description Framework, or RDF, which is a data model for representing metadata [W3C, 2004a]. With RDF, metadata is represented in *triples*, each consisting of a subject, a predicate, and an object. Sometimes a triple is also called a *statement*.

RDF is powerful in terms of representing complex metadata and relationships. For instance, a book sold online can be thought of a creative work and a product at the same time. With BigSemantics, it is currently difficult to represent metadata instances of multiple types. With RDF, let `B` be the book sold online, `rdf:type` be the predicate indicating "<subject> is an instance of type <object>", and `CW` and `P` be the types of creative work and product, respectively, the above example can be represented by two triples: `B rdf:type CW` and `B rdf:type P`. Another powerful representation RDF supports is *reification*, which enables referring to another statement (rather than a resource). Reification can be useful for representing "second order" metadata, such as provenance and credibility.

RDF is primarily used to represent metadata. The Semantic Web approach also provides pow-

erful tools for describing metadata schemas. RDFS is a data model built on top of RDF, and can be used to describe schemas for metadata represented in RDF [W3C, 2004b]. RDFS provides conceptual constructs such as classes and properties, as tools for describing data models (a.k.a. schemas) used in user tasks or applications. The Web Ontology Language, or OWL, provides even more tools for describing schemas, such as union types, constraints, and annotations [W3C, 2009]. The goal of OWL is not just to describe schemas, but also to support computations and inferences on the information it describes. In other words, OWL describes *ontologies*, which are formal representations of *knowledge* in a domain.

Besides expressive standards such as RDF, RDFS, and OWL, the Semantic Web approach also provides powerful technologies, such as triple stores and structured query engines. If information on the web is represented, organized, and used with these standards and technologies, scenarios imagined by the Semantic Web pioneers would come true. However, one major barrier to realizing the Semantic Web dream is that not many websites publish their data in RDF / RDFS / OWL. At the end, websites are built for humans to use, not for machines. The dominant technologies for the web—HTML, CSS, and JavaScript—are designed with a great focus on visual presentation and interactivity, rather than representing formal, abstract, and machine readable data models. Webmasters are more concerned about if their website satisfies human users' needs, than about publishing their data in RDF. RDF's serialized formats, such as RDF/XML, are usually verbose, making it difficult for humans to read and use. The result is a dearth of RDF described resources available on the web. Today, despite some useful datasets, such as Wikipedia infoboxes, are available in RDF triples, the Semantic Web approach is yet to be widely adopted [Lytras and García González, 2008]. Even when RDF triples are available, researchers have criticized the approach in multiple aspects [Marshall and Shipman, 2003]. For example, due to the situated nature of knowledge, what is provided in RDF triples may not be what the user needs in a particular contexts; this is likely to happen at some point of time because it is difficult to anticipate all the possible uses of data. It is reasonable to expect that maintaining consistent schemas across websites over an extended period of time to be difficult and expensive,

More recent efforts on web standardization attempt to make it easier for websites to publish formal metadata. For example, microdata is designed to be embedded in HTML, in the form of special attributes on regular HTML tags [W3C, 2012]. Microdata is supported by major search engines and other popular online services, so that webmasters can use it to optimize their presence and presentation on the web. Advocates also put together a large set of commonly used schemas, at `schema.org`, to facilitate sharing. While microdata is much easier to publish and use than typical RDF, users are still limited by what is published by the website.

The meta-metadata language provided by BigSemantics is not as expressive or powerful as the Semantic Web standards. However, our approach with BigSemantics and dynamic exploratory browsing interfaces addresses gaps in the Semantic Web standards or technologies:

1. BigSemantics supports client-side metadata modeling and extraction, thus does not rely on websites to publish data in RDF, and allows application developers to decide what metadata to use and how;

2. Dynamic exploratory browsing interfaces, such as MICE, presents extracted metadata in useful, interactive forms.

Support for client-side metadata modeling and extraction is based on the situatedness of human tasks: webmasters may not be able to anticipate every possible way their website can be used in applications, thus it is better to let application developers decide. Addressing presenting metadata in interactive forms is identified by previous research as an important challenge [Bizer et al., 2011], and is essential for making metadata useful for end users. We hope that as more and more users see the value of (appropriately presented) metadata in their tasks, more and more developers will build usable applications operating on these metadata types and datasets, and eventually more and more websites will publish useful metadata.

6.3 Limitations and Future Work

The interdisciplinary model of scholarly research ideation is meant to inform the design of: studies of scholarly research practice; interactive systems that support scholarly research practice;

evaluations of interactive systems that support scholarly research practice; and education in scholarly fields. Future work is needed to more specifically study each component of the model, and their relationships, in the context of scholarly research ideation. The model itself is limited by the scope of the present study of engagement with scholarly material by fairly established researchers in fields around human-computer interaction. Future research could beneficially investigate how these findings—and our associated model—are general, and how specificities arise, based on participants' fields and other contextual characteristics. Future research would benefit from examining the roles of material beyond prior work, such as raw data collected through studies, in research ideation. More research is needed to test and refine the model with participants from other research fields and in other information intensive processes.

The present research focuses on mini-c creativity (i.e. personally meaningful experiences and interpretations) experienced by investigators in scholarly research, because these mini-c steps are often essential for developing ideas of larger impact. Investigating how scholars collaborate in developing novel research, and how the model can be applied in creative tasks of larger impact, is an interesting future direction.

Besides engaging with prior work, researchers generate new ideas from a wide range of activities, such as analyzing raw data and material [Caro, 2019] and discussing with colleagues. Future studies on scholarly research ideation can investigate how people engage in those activities, and integrate or refine our model with new findings.

Processes addressed by the model are by no means limited to scholarly research ideation. The model has the potential to describe diverse creative tasks, ranging from planning a weekend to writing an essay to designing a campaign, all of which involve seeking information, and putting found information together in new, situated contexts, to support developing ideas. Previous research refers to these as *information-based ideation* tasks [Kerne et al., 2014]. We argue that the new model is likely generalizable, to understanding practices of and designing experiences for diverse information-based ideation tasks. Further research can work to identify what is common and what changes in particular ideation activity contexts, and how these findings can help us under-

stand and support innovation in diverse fields. Since creative innovation is so essential to personal satisfaction, sustainability, and economic success [Boden, 2004; Kaufman and Beghetto, 2009; National Academy of Engineering, 2005], the ideation model's potential impact is extensive.

Dynamic interfaces based on the metadata type system have the potential to transform browsing experiences with web information for a wide range of open-ended, exploratory tasks. Exploratory browsing interfaces can be embedded into HTML pages to transform passive hyperlinks, enriching diverse and integral information experiences, including digital libraries, shopping, social networks, messaging services, email clients, and newspapers. Our open source implementations of the type system and MICE have the potential to facilitate the engagement of research, open source, and industry communities in engineering new interactive systems in diverse domains for exploratory browsing and search. The metadata type system enables a new family of dynamic interfaces that help users browse the WWW. Support for exploratory browsing while maintaining context will be valuable in sensemaking and berrypicking tasks. Future research can incorporate these techniques, to develop new support for exploratory search, and integrate it with exploratory browsing.

The BigSemantics metadata type system is currently limited in terms of representing complex structures and relationships. It is difficult to represent instances of two or more types. When websites change their design, wrappers need to be manually updated. The current MICE interface is designed for general metadata; more specific designs are needed for specific use cases. Features such as sorting and filtering by certain metadata fields and dimensions (aka *facets*) in MICE are wanted by study participants. Future work should address these limitations and feature requests.

7. CONCLUSION

Semantic information (a.k.a. metadata), when presented in usable forms as in MICE, supports scholarly research ideation through:

- Enabling citation chaining and other forms of exploratory browsing, thus helping researchers find more interesting prior work, which is crucial for staying aware of the field and developing new contributions;
- Providing useful contextual information in researchers' prior work curation process and product, thus facilitating reflecting, developing relationships, and generating new ideas.

Today, as most researchers heavily rely on the Internet and digital libraries for browsing and curating prior work, web semantics plays an active and important role in supporting scholarly research ideation.

Performing scholarly research ideation tasks often entails thinking about multiple ideas at the same time, which are often remotely related, sometimes conflicting, and synthesizing them into a cohesive one. It therefore faces the fundamental limitations of human cognition: people's working memory is limited to about 4 chunks at a time [Cowan, 2001]. This mandates that interfaces supporting scholarly research ideation need to present information in ways that help users overcome this limitation and effectively use their limited cognitive resources. Challenges include:

- Present as much information as needed by the ideation task, without causing information overload;
- Keep it simple and easy to understand, without losing structure and intricacy;
- Support as many sources as possible, without adding complexity;
- Support organization of information elements into chunks, as needed (potentially recursively);

- Promote exploration and discovery, without losing context.

MICE addresses the above requirements by presenting generic web semantics in innovative ways. MICE uses the same generic `scholarly_article` schema for different underlying information sources. This abstraction helps remove unnecessary details and complexities from the ways different information sources model their semantic information, so that users can focus on the activities that lead to ideation, such as citation chaining, exploring, connecting, and comparing. Web semantics are organized with hierarchical fields. In MICE, composite and collection fields can be collapsed, to save space and user's attention, or expanded, to afford viewing details. This simple presentation is intuitive to use, and at the same time enables users to flexibly control the level of details they would like to see and process at that moment. MICE's ability to dynamically retrieve and present semantic information from multiple sources in the same context (with relationships visible) addresses the need for citation chaining while maintaining context. Overall, MICE, and potentially other dynamic exploratory browsing interfaces, can be used to effective support scholarly research ideation tasks.

Web semantics is useful in curation—an art-based concept and method for putting together a transformative context consisting of found objects and materials—which we found valuable for characterizing essential scholarly research ideation practices. Every participant researcher in our study engages in curating prior work, which they find interesting. In curation processes, they assemble a new context consisting of found work, along with their own interpretations, critiques, and reflections, as a method and medium for developing new ideas. This kind of contextualizing is essential for conducting scholarly research: having ideas about what interesting prior work is out there and how it relates to their own research is a precondition for deriving and demonstrating novel and valuable contributions. When it comes to processes and products of curation, external representations become important: our embodied minds combine perceptual, emotional, and cognitive capabilities when dealing with complex tasks such as ideation. External representations facilitate such combination. Usable presentations of web semantics, such as those in MICE, function as external representations in curation, providing structured contextual information in simple

and concise forms.

This research is grounded in a perspective, in which we see creative ideation as the motivation and central raison d'être of scholarly research. After all, scholarly research is about creating new, validated ideas. Notably, we take a mini-c approach, focusing our investigation into researchers' personal, yet meaningful and transformative experiences of new encounters and interpretations, such as finding an interesting citation and discovering unexpected connections. The shift of focus towards mini-c ideation is because first, in many cases, eminent innovation grows out of personal creative ideation, and second, in order to promote creative ideation, we need to study its genesis. We believe that this approach will be valuable for investigations into creative ideation tasks in other contexts, too.

Applying this ideation-centric, mini-c approach to our study with researchers' practices, and drawing on conclusions from information science, we develop the interdisciplinary model of scholarly research ideation. In creative cognition, models "can provide a deeper understanding of how creativity is expressed across widely varied domains" [Finke et al., 1992]. As an example, the Geneplore model was introduced to "provide a possible foundation for a unified account of creative cognition" [Finke et al., 1992]. The present, interdisciplinary model of scholarly research ideation incorporates Geneplore, and goes beyond, by integrating human processes involving both internal and external representations, focusing on both learning and creating. Information foraging is an example of a prior model, whose form is a bidirectional chain consisted of many loops [Pirolli and Card, 2005]. The present interdisciplinary model of scholarly research model uses symmetric, full interconnections among components, in order to represent the integrated and situated nature of the underlying human processes: after all, the brain and the connected neural system works as an interconnected and coordinated network [Beaty et al., 2016; McIntosh, 2000]. We expect this model to be useful in guiding the design of new scholarly-research-supporting interfaces, as well as investigations into researchers' practices and experiences.

Under the hood, borrowing object-oriented programming concepts and constructs—such as inheritance, polymorphism, and dynamic dispatch—this research develops a novel approach to en-

gineering usable exploratory browsing interfaces, for working with heterogeneous metadata, such as MICE. Types integrate metadata data models, extraction, and presentation. Seemingly contrary to common practices of separating concerns, this integration is demanded by how these aspects of exploratory browsing are inherently connected in vital user experiences. Integrative metadata types operationalize dynamic exploratory browsing interfaces by enabling recursive extraction and usable presentation of heterogeneous metadata summaries from diverse sources.

Metadata summary representations produced by the type system enable reduced, yet expandable presentation of web pages. Presentation semantics enable the user to browse original web pages, as needed. Study participants found that MICE helped reduce disorientation and digression by displaying metadata in one context and making relationships visible, including to previously encountered information.

The dynamic nature of such interfaces is essential to exploratory browsing. When the user serendipitously seeks to explore new information encountered through links, the interface dynamically expands, using types to customize metadata derivation and presentation. Further, newly published information can be dynamically incorporated. Thus, the metadata type system fundamentally differs from technologies that only operate on datasets assembled in advance.

The custom presentation semantics specified in types, such as ordering, formatting, hiding, and hyperlinking field values, enable type-specific emphasis that can mitigate the cognitive load inherent in browsing large amounts of information. The type system and MICE constitute a practical method for building web-scale dynamic exploratory interfaces. Study participants found MICE's concise presentation of linked metadata usable and valuable for exploratory browsing.

Creative innovation has emerged as a crucial factor for personal wellbeing, growth, and national economic success [Boden, 2004; Kaufman and Beghetto, 2009; National Academy of Engineering, 2005]. Ideation processes are neither entirely logical nor completely random. They are rooted in serendipity, depend on emergent phenomena that arise in learning and creating, and require iterative, interwoven cycles of sensemaking, seeking, curation, generating, and exploring. The present research connects strengths from different areas: understandings of human cognitive

processes, findings on people's practices working with information, art practices, knowledge on designing and engineering dynamic, usable presentations and visualizations of information, and technologies of obtaining abundant and rich information on the web. The result is that we take new steps toward an interdisciplinary theory of researchers' tasks and practices as a whole, and new interfaces supporting scholarly research ideation, an important creative task. We believe that this interdisciplinary approach can be applied to investigations into other fields where information-based ideation plays an essential role.

REFERENCES

Abbott, D. (2008). Dcc briefing paper: What is digital curation?

Adler, M. J. and van Doren, C. (1972). *How To Read A Book: The Classic Guide To Intelligent Reading*. Simon and Schuster.

AlNoamany, Y., Weigle, M. C., and Nelson, M. L. (2017). Generating stories from archived collections. In *Proceedings of ACM WebSci*, pages 309–318.

Ankolekar, A. et al. (2007). The two cultures: Mashing up web 2.0 and the semantic web. In *Proceedings of WWW*, pages 825–834.

Antoniou, G. and van Harmelen, F. (2004). *A Semantic Web Primer*. The MIT Press.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13(5):407–424.

Bateson, G. (1972). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. University of Chicago Press.

Beaty, R. E., Benedek, M., Silvia, P. J., and Schacter, D. L. (2016). Creative cognition and brain network dynamics. *Trends in Cognitive Sciences*, 20(2):87–95.

Beghetto, R. A. and Kaufman, J. C. (2007). Toward a broader conception of creativity: A case for "mini-c" creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 1(2):73.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information and Library Science*, 5:133–143.

Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). ASK for information retrieval: Part I. background and theory. *Journal of Documentation*, 38(2):61–71.

Berners-Lee, T. (1994). RFC 1738: Uniform resource locators (URL). *RFC*.

Berners-Lee, T. et al. (2006). Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of SWUI*.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):34–43.

Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227. IGI Global.

Blandford, A. and Attfield, S. (2010). Interacting with information. *Synthesis Lectures on Human-Centered Informatics*, 3(1):1–99.

Boardman, R. and Sasse, M. A. (2004). Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *Proceedings of ACM CHI*, pages 583–590.

Boden, M. A. (2004). *The Creative Mind: Myths and Mechanisms*. Psychology Press.

Booch, G., Maksimchuk, R., Engle, M., Young, B., Conallen, J., and Houston, K. (2007). *Object-Oriented Analysis and Design with Applications*. Pearson Education.

Bookstein, A. and Cooper, W. (1976). A general mathematical model for information retrieval systems. *The Library Quarterly*, 46(2):153–167.

Bowker, G. C. and Star, S. L. (2000). *Sorting Things Out: Classification and Its Consequences*. MIT press.

Bracha, G. and Cook, W. (1990). Mixin-based inheritance. In *Proceedings of OOPSLA/ECOOP*.

Breton, A. (1924). Manifesto of surrealism. *Manifestoes of Surrealism*, 15.

Brewer, W. F. (1987). Schemas versus mental models in human memory. In Morris, P., editor, *Modeling Cognition*, pages 187–197. John Wiley & Sons.

Bush, V. (1945). As we may think. *Atlantic Monthly*, 176.

Caro, R. A. (2019). The secrets of Lyndon Johnson's archives. *The New Yorker*.

Center for History and New Media at George Mason University (2006). Zotero. `http://www.zotero.org`. Visited Sept 24, 2015.

Charmaz, K. (2006). *Constructing Grounded Theory*. Sage Publications.

Chi, E. H., Pirolli, P., Chen, K., and Pitkow, J. (2001). Using information scent to model user information needs and actions and the web. In *Proceedings of ACM CHI*, pages 490–497.

Clark, T. (2015). *The Painting of Modern Life: Paris in the Art of Manet and His Followers*. Princeton University Press.

Conklin, J. (1987). Hypertext: an introduction and survey. *Computer*, 20(9):17–41.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, 24(1):87–114.

Cutting, D. R. et al. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of ACM SIGIR*.

Dallas, C. (2016). Digital curation beyond the "wild frontier": a pragmatic approach. *Archival Science*, 16(4):421–457.

Deligiannidis, L. et al. (2007). RDF data exploration and visualization. In *Proceedings of CIMS*, pages 39–46. ACM.

Dontcheva, M., Drucker, S. M., Wade, G., Salesin, D., and Cohen, M. F. (2006). Summarizing personal web browsing sessions. *Proceedings of ACM UIST*, pages 115–124.

Dontcheva, M. et al. (2008). Experiences with content extraction from the web. In *Proceedings of ACM UIST*.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1):19–30.

Edwards, D. M. and Hardman, L. (1999). Lost in hyperspace: cognitive mapping and navigation in a hypertext environment. In *Hypertext: Theory into Practice*, pages 90–105. Intellect Books, Exeter, UK.

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, 45(3):171–212.

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

Finke, R. A., Ward, T. B., and Smith, S. M. (1992). *Creative Cognition: Theory, Research, and Applications*. MIT Press, Cambridge, MA.

Foss, C. L. (1989). Detecting lost users: Empirical studies on browsing hypertext. Technical report.

Foster, A. and Ford, N. (2003). Serendipity and information seeking: An empirical study. *Journal*

of Documentation, 59(3):321–340.

Gamow, G. (1985). *Thirty Years that Shook Physics: The Story of Quantum Theory*. Courier Corporation.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479.

Garfield, E. (1995). New international professional society signals the maturing of scientometrics and informetrics. *The Scientist*, 9(16):11.

Gaver, W. W., Beaver, J., and Benford, S. (2003). Ambiguity as a resource for design. In *Proceedings of ACM CHI*, pages 233–240.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Gentner, D. and Gentner, D. R. (1982). Flowing waters or teeming crowds: Mental models of electricity. Technical report, DTIC Document.

Giaccardi, E. and Karana, E. (2015). Foundations of materials experience: An approach for HCI. In *Proceedings of ACM CHI*, pages 2447–2456.

Glaser, H. et al. (2004). CS AKTive Space: Building a semantic web application. In *Proceedings of ESWS*, pages 417–432. Springer Verlag.

Glenberg, A. M. (2015). Few believe the world is flat: How embodiment is changing the scientific understanding of cognition. *Canadian Journal of Experimental Psychology*, 69(2):165–71.

Glenberg, A. M. and Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9(3):558–565.

Glenberg, A. M. and Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language*, 31(2):129–151.

Goldschmidt, G. (1991). The dialectics of sketching. *Creativity Research Journal*, 4(2):123–143.

Gray, J., Szalay, A. S., Thakar, A. R., Stoughton, C., and vandenBerg, J. (2002). Online scientific data curation, publication, and archiving. In *Proceedings of SPIE*, volume 4846, pages 103–107.

Greenberg, S. and Cockburn, A. (2002). Getting back to back: Alternate behaviors for a web browser's back button. In *Proceedings of HFWEB*.

Hearst, M. A. and Pedersen, J. O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of ACM SIGIR*.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, 3(1):134–140.

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM ToCHI*, 7(2):174–196.

Hutchins, E. (1995). *Cognition in the Wild*. A Bradford book. MIT Press.

Huynh, D. F. and Karger, D. (2009). Parallax and companion: set-based browsing for the data web. In *Proceedings of WWW*. ACM.

Huynh, D. F., Karger, D. R., and Miller, R. C. (2007). Exhibit: Lightweight structured data publishing. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 737–746. ACM.

Huynh, D. F., Mazzocchi, S., and Karger, D. (2005). Piggy Bank: Experience the semantic web inside your web browser. In *Proceedings of ISWC*.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, 20(3):341–367.

Interface Ecology Lab (2014). Metadata In-Context Expander (MICE) Demo. `http://ecologylab.net/mice`.

Interface Ecology Lab (2016). An open source metadata type system implementation. `https://github.com/ecologylab/BigSemantics/wiki`.

International DOI Foundation (2018). The digital object identifier system. `http://www.doi.org/`.

Jain, A., Lupfer, N., Qu, Y., Linder, R., Kerne, A., and Smith, S. M. (2015). Evaluating tweetbubble with ideation metrics of exploratory browsing. In *Proceedings of ACM Creativity & Cognition*, pages 53–62.

Järvelin, K. and Ingwersen, P. (2004). Information seeking research needs extension towards tasks and technology. *Information Research*, 10(1).

Jones, W. (2007). Personal information management. *Annual Review of Information Science and Technology*, 41(1):453–504.

Jones, W., Bruce, H., and Dumais, S. (2001). Keeping found things found on the web. In *Proceedings of ACM CIKM*, pages 119–126.

Kaufman, J. C. and Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity. *Review of General Psychology*, 13(1):1–12.

Kerne, A. (2001). *CollageMachine: A Model of "Interface Ecology"*. PhD thesis, New York University.

Kerne, A., Lupfer, N., Linder, R., Qu, Y., Valdez, A., Jain, A., Keith, K., Carrasco, M., Vanegas, J., and Billingsley, A. (2017). Strategies of free-form web curation: Processes of creative engagement with prior work. In *Proceedings of ACM Creativity & Cognition*.

Kerne, A., Qu, Y., Webb, A. M., Damaraju, S., Lupfer, N., and Mathur, A. (2010). Meta-Metadata: A metadata semantics language for collection representation applications. In *Proceedings of ACM CIKM*, pages 1129–1138. ACM.

Kerne, A. and Smith, S. M. (2004). The information discovery framework. In *Proceedings of ACM DIS*, pages 357–360.

Kerne, A., Webb, A. M., Smith, S. M., Linder, R., Lupfer, N., Qu, Y., Moeller, J., and Damaraju, S. (2014). Using metrics of curation to evaluate information-based ideation. *ACM ToCHI*, 21(3):14:1–14:48.

Kiewra, K. A. (1989). A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review*, 1(2):147–172.

Kirsh, D. (2010). Thinking with external representations. *AI & Society*, 25(4):441–454.

Klemmer, S. R. et al. (2002). Where do web sites come from?: capturing and interacting with design history. In *Proceedings of ACM CHI*, pages 1–8.

Kolko, J. (2007). Information architecture and design strategy: The importance of synthesis during

the process of design. In *Industrial Designers Society of America Conference*, pages 17–20.

Kolko, J. (2010). Abductive thinking and sensemaking: The drivers of design synthesis. *Design Issues*, 26(1):15–28.

Krauss, R. E. (1994). *The Optical Unconscious*. MIT Press.

Lakoff, G. and Johnson, M. (1999). *Philosophy In The Flesh*. Collection of Jamie and Michael Kassler. Basic Books.

Landow, G. P. (2006). *Hypertext 3.0: Critical Theory and New Media in an Era of Globalization*. JHU Press.

Lin, J. et al. (2009). End-user programming of mashups with vegemite. In *Proceedings of IUI*, pages 97–106. ACM.

Linder, R., Snodgrass, C., and Kerne, A. (2014). Everyday ideation: All of my ideas are on pinterest. In *Proceedings of ACM CHI*, pages 2411–2420.

Linsey, J. S., Wood, K., and Markman, A. (2008). Increasing innovation: Presentation and evaluation of the wordtree design-by-analogy method. In *Proceedings of ASME 2008 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, pages 21–32. American Society of Mechanical Engineers.

Lippard, L. R. (1971). *Dadas on Art*. Prentice Hall.

Lippard, L. R. (1973). *Six Years: The Dematerialization of the Art Object from 1966 to 1972*. University of California Press.

Lytras, M. D. and García González, R. (2008). Semantic web applications: A framework for industry and business exploitation–what is needed for the adoption of the semantic web from the market and industry. *International Journal of Knowledge and Learning*, 4(1):93–108.

Marchionini, G. (1997). *Information Seeking in Electronic Environments*. Number 9 in Cambridge Series on Human-Computer Interaction. Cambridge University Press.

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of ACM*, 49(4):41.

Marchionini, G. and Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext

systems. *Computer*, 21(1):70–80.

Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of ACM HYPERTEXT*, pages 31–40.

Marshall, C. C. (1997). Annotation: From paper books to the digital library. In *Proceedings of ACM DL*, pages 131–140.

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In *Proceedings of ACM HYPERTEXT*, pages 40–49.

Marshall, C. C. (2008). From writing and analysis to the repository: Taking the scholars' perspective on scholarly archiving. In *Proceedings of ACM/IEEE JCDL*, pages 251–260.

Marshall, C. C. and Bly, S. (2005). Saving and using encountered information: Implications for electronic periodicals. In *Proceedings of ACM CHI*, pages 111–120.

Marshall, C. C., Price, M. N., Golovchinsky, G., and Schilit, B. N. (1999). Introducing a digital library reading appliance into a reading group. In *Proceedings of ACM DL*, pages 77–84.

Marshall, C. C. and Shipman, F. M. (2003). Which semantic web? In *Proceedings of ACM HYPERTEXT*.

Marshall, C. C. and Shipman, III, F. M. (1995). Spatial hypertext: Designing for change. *Communications of ACM*, 38(8):88–97.

McIntosh, A. R. (2000). Towards a network theory of cognition. *Neural Networks*, 13(8-9):861–870.

Merleau-Ponty, M. (1962). *Phenomenology of Perception*. Routledge.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(1):81–97.

Miller, P. D. (2004). *Rhythm Science*. MIT Press.

Mitchell, J. C., John, C., and Apt, K. (2003). *Concepts in Programming Languages*. Cambridge University Press.

National Academy of Engineering (2005). *Engineering Research and America's Future: Meeting the Challenges of a Global Economy*. The National Academies Press.

Nersessian, N. J. (1999). Model-based reasoning in conceptual change. In *Model-Based Reasoning in Scientific Discovery*, pages 5–22. Springer US, Boston, MA.

Neuwirth, C. M. and Kaufer, D. S. (1989). The role of external representation in the writing process: Implications for the design of hypertext-based writing tools. In *Proceedings of ACM Hypertext*, pages 319–341.

Norman, D. (2002). *The Design of Everyday Things*. Basic Books.

OED (2017). Oxford English Dictionary. Accessed April 26, 2017.

Olby, R. (2013). *The Path to the Double Helix: The Discovery of DNA*. Courier Corporation.

O'Neill, P. (2012). *The Culture of Curating and the Curating of Culture(s)*. MIT Press.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Pham, H. et al. (2012). Clui: A platform for handles to rich objects. In *Proceedings of ACM UIST*, pages 177–188. ACM.

Pierce, B. C. and Benjamin, C. (2002). *Types and Programming Languages*. MIT press.

Pirolli, P. and Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, volume 5, pages 2–4.

Pirolli, P., Schank, P., Hearst, M., and Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–220. ACM.

Qu, Y., Kerne, A., Lupfer, N., Linder, R., and Jain, A. (2014). Metadata type system: Integrate presentation, data models and extraction to enable exploratory browsing interfaces. In *Proceedings of ACM Engineering Interactive Computing Systems*, pages 107–116.

Rástočný, K. et al. (2011). Supporting search result browsing and exploration via cluster-based views and zoom-based navigation. In *Proceedings of WI-IAT*, volume 3.

Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of ACM INTERACT and CHI*, pages 269–276.

Samp, K. et al. (2008). Atom interface: A novel interface for exploring and browsing semantic space. In *Proceedings of SWUI at CHI*.

Scaife, M. and Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, 45(2):185–213.

Schilit, B. N., Golovchinsky, G., and Price, M. N. (1998). Beyond paper: Supporting active reading with free form digital ink annotations. In *Proceedings of ACM CHI*, pages 249–256.

Schön, D. A. (1992). Designing as reflective conversation with the materials of a design situation. *Knowledge-Based Systems*, 5(1):3–14.

schraefel, m. c., Karam, M., and Zhao, S. (2003). mSpace: interaction design for user-determined, adaptable domain exploration in hypermedia. In *Proceedings of AH: Workshop on Adaptive Hypermedia and Adaptive Web Based Systems*.

Science History Institute (2017). James Watson, Francis Crick, Maurice Wilkins, and Rosalind Franklin. `https://www.sciencehistory.org/historical-profile/james-watson-francis-crick-maurice-wilkins-and-rosalind-franklin`.

Seitz, W. C. (1961). *The Art of Assemblage*. Museum of Modern Art (New York).

Sellen, A. J. and Harper, R. H. (2003). *The Myth of the Paperless Office*. MIT Press.

Shah, J. J. (1998). Experimental investigation of collaborative techniques for progressive idea generation. In *Proceedings of ASME DETC Design Theory and Methodology Conference*, pages 13–16.

Shipman, III, F. M. and Marshall, C. C. (1999). Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work (CSCW)*, 8(4):333–352.

Simon, H. A. (1971). Designing organizations for an information-rich world. *Computers, Communications, and the Public Interest*, 72:37.

Smith, D. A. (2011). *Exploratory and Faceted Browsing, over Heterogeneous and Cross-Domain Data Sources*. PhD thesis, U of Southampton.

Smith, S. M. and Blankenship, S. E. (1991). Incubation and the persistence of fixation in problem

solving. *The American Journal of Psychology*, 104(1):61–87.

Star, S. L. and Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. *Social Studies of Science*, 19(3):387–420.

Stegemann, T. et al. (2012). Interactive construction of semantic widgets for visualizing semantic web data. In *Proceedings of ACM Engineering Interactive Computing Systems*, pages 157–162.

Stoan, S. K. (1984). Research and library skills: An analysis and interpretation. *College and Research Libraries*, 45(2):99–109.

Suwa, M. and Tversky, B. (2002). External representations contribute to the dynamic construction of ideas. In *Proceedings of Diagrammatic Representation and Inference*, pages 341–343. Springer Berlin Heidelberg.

Thomas, J. J. and Cook, K. (2006). A visual analytics agenda. *Computer Graphics and Applications*, 26(1):10–13.

Thorne, K. (1994). *Black Holes & Time Warps: Einstein's Outrageous Legacy*. W. W. Norton & Company.

Tufte, E. (1990). *Envisioning Information*. Graphics Press.

Tversky, B. (1993). Cognitive maps, cognitive collages, and spatial mental models. In Frank, A. U. and Campari, I., editors, *Proceedings of COSIT: Spatial Information Theory A Theoretical Basis for GIS*, pages 14–24. Springer Berlin Heidelberg.

Varela, F., Thompson, E., and Rosch, E. (1992). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

W3C (2004a). RDF primer. `http://www.w3.org/TR/2004/REC-rdf-primer-20040210/`.

W3C (2004b). RDF vocabulary description language 1.0: RDF schema. `http://www.w3.org/TR/rdf-schema/`.

W3C (2009). OWL2 web ontology language document overview. `http://www.w3.org/TR/owl2-overview/`.

W3C (2010). XML path language. https://www.w3.org/TR/xpath.

W3C (2011). Cascading style sheets level 2 revision 1 (CSS 2.1) specification. https://www.w3.org/TR/2011/REC-CSS2-20110607/.

W3C (2012). HTML5: A vocabulary and associated apis for html and xhtml. http://www.w3.org/TR/html5/microdata.html.

Webb, A. M., Kerne, A., Linder, R., Lupfer, N., Qu, Y., Keith, K., Carrasco, M., and Chen, Y. (2016). A free-form medium for curating the digital. In *Curating the Digital*, pages 73–87. Springer.

Webb, A. M., Linder, R., Kerne, A., Lupfer, N., Qu, Y., Poffenberger, B., and Revia, C. (2013). Promoting reflection and interpretation in education: Curating rich bookmarks as information composition. In *Proceedings of ACM Creativity & Cognition*, pages 53–62.

White, R. W., Kules, B., Drucker, S. M., and schraefel, m. (2006a). Introduction to supporting exploratory search. *Communications of ACM*, 49(4):36–39.

White, R. W., Kules, B., Drucker, S. M., and schraefel, m. (2006b). Supporting exploratory search, intro to special issue. *Communications of ACM*, 49(4).

White, R. W. and Roth, R. A. (2009). Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98.

Whittaker, S. (2011). Personal information management: from information consumption to curation. *Annual Review of Information Science and Technology*, 45(1):1–62.

Wiener, N. (1961). *Cybernetics Or Control and Communication in the Animal and the Machine*. MIT Press.

Wikipedia editors (2018). Exploratory search. http://en.wikipedia.org/wiki/Exploratory_search.

Wilkenfeld, M. J. and Ward, T. B. (2001). Similarity and emergence in conceptual combination. *Journal of Memory and Language*, 45(1):21–38.

Wilson, M. L., Kules, B., m. c. schraefel, and Shneiderman, B. (2010). From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web*

Science, 2(1):1–97.

Winograd, T. and Flores, F. (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Intellect Books.

Wong, J. and Hong, J. I. (2007). Making mashups with Marmite: Towards end-user programming for the web. In *Proceedings of ACM CHI*. ACM.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, 21(2):179–217.

Zhang, J. and Norman, D. A. (1995). A representational analysis of numeration systems. *Cognition*, 57(3):271–295.

Zhao, X. and Lindley, S. E. (2014). Curation through use: Understanding the personal value of social media. In *Proceedings of ACM CHI*, pages 2431–2440.