REPORTING SERENDIPITY IN BIOMEDICAL RESEARCH LITERATURE:

A MIXED-METHODS ANALYSIS

A Dissertation

Presented To

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Carla McCaghren Allen

Sanda Erdelez, PhD, Dissertation Supervisor

Chi-Ren Shyu, PhD, Dissertation Supervisor

December, 2018

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

REPORTING SERENDIPITY IN BIOMEDICAL RESEARCH LITERATURE:

A MIXED-METHODS ANALYSIS

presented by Carla M. Allen,

a candidate for the degree of doctor of philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

_____

Professor Sanda Erdelez

_____

Professor Chi-Ren Shyu

_____

Professor Denice Adkins

_____

Professor Terry Barnes

To my Creator, Lord and Savior,

Who gives me the strength to do all things…

Who sees the end from the beginning….

And has a purpose for everything that happens,

Even those things that we perceive as serendipitous.

Acknowledgements

To use a favorite quote from the founding Dean of the School of Health Professions at the University of Missouri, Richard Oliver, "If you see a turtle on a fence post, you know he didn't get there by himself." The process of completing a doctoral program has left me feeling like a turtle on a fence post. Without the amazing support and encouragement of those around me, I could never have completed this stage of the journey.

First, I would like to express my deep gratitude to the members of my doctoral committee. I greatly appreciate my academic advisor and the chair of my dissertation committee, Dr. Sanda Erdelez, for her academic enthusiasm, patient guidance, and continuous encouragement throughout my study at the University of Missouri. Our journey together was serendipitous. With what started as a routine introduction in an online class, I found myself involved in a research team that will forever shape the way I approach research endeavors. I would also like to recognize Dr. Chi-Ren Shyu, the PI of that project and co-chair of my dissertation committee. As a part of that team, I was able to observe the mentoring styles of both Dr. Erdelez and Dr. Shyu as they guided other doctoral candidates through the completion of their dissertation studies. I witnessed the way Dr. Shyu coordinated our research team, insuring that everyone on the team brought their unique disciplinary perspectives to the projects and gained something valuable from the experience. I hope to be able to emulate that process with my own doctoral students someday. I appreciate the depth and richness of cultural experience that Dr. Denice Adkins showed me was so critical to information experiences, and the need to consider context when designing and analyzing information systems. Finally, I wish to thank Dr.

Terry Barnes who has mentored my professional growth from the beginning of my graduate studies. He taught me to build on ideas from simplest to most complex and to focus on meeting the needs of society as a whole when designing instruction or research.

In addition to my committee, there are a host of people who have significantly contributed to my success:

To Adria Koehn, my sister from a different mother who is teaching me about boundaries and work-life balance, and without whom the Table of Figures would never have gotten formatted!

To Dr. Osasuyi Dirisu, for her statistical insight and support.

To the doctoral students who allowed me to learn research methods alongside them and gain perspective on my own work as a result: Dr. Weichao Chen, Dr. Xin Wang, Dr. Hongfei Cao, Dr. Dinara Saparova, and Blake Anderson.

To my professional mentors and colleagues: Dr. Glen Heggie, Dr. Shawna Strickland, Pat Tew, Mary Sebacher, Sharlette Anderson, Dr. Kathy Moss, Jenny Keeley, and Sara Parker. Thanks for standing by me and encouraging me.

To the SHP researchers who started me on this path: Dr. Erin Danneker and Dr. Judith Goodman.

Finally, I am deeply indebted to my family. Without their understanding and love, I would never have completed this journey.

Table of Contents

Table of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Table of Figures

Chapter 2

Chapter 3

Abstract

As serendipity is an unexpected, anomalous, or inconsistent observation that culminates in a valuable, positive outcome (McCay-Peet & Toms, 2018, pp. 4–6), it can be inferred that effectively supporting serendipity will result in a greater incidence of the desired positive outcomes (McCay-Peet & Toms, 2018, p. 22). In order to effectively support serendipity, however, we must first understand the overall process or experience of serendipity and the factors influencing its attainment. Currently, our understanding and models of the serendipitous experience are based almost exclusively on example collections, compilations of examples of serendipity that authors and researchers have collected as they encounter them (Gries, 2009, p. 9). Unfortunately, reliance on such collections can lead to an over-representation of more vivid and dramatic examples and possible under-representation of more common, but less noticeable, exemplars. By applying the principles of corpus research, which involves electronic compilation of examples in existing documents, we can alleviate this problem and obtain a more balanced and representative understanding of serendipitous experiences (Gries, 2009). This three-article dissertation describes the phenomenon of serendipity, as it is recorded in biomedical research articles indexed in the PubMed Central database, in a way that might inform the development of machine compilation systems for the support of serendipity. Within this study, serendipity is generally defined as a process or experience that begins with encountering some type of information. That information is subsequently analyzed and further pursued by an individual with related knowledge, skills, and understanding, and, finally, allows them to realize a valuable outcome. The information encounter that initiates the serendipity

experience exhibits qualities of unexpectedness as well as value for the user. In this mixed method study, qualitative content analysis, supported by natural language processing, and concurrent with statistical analysis, is applied to gain a robust understanding of the phenomenon of serendipity that may reveal features of serendipitous experience useful to the development of recommender system algorithms.

Keywords

serendipity; information behavior; content analysis; information encountering; user experience; human-centered information retrieval; research reporting; recommender systems.

Chapter 1 Introduction

Problem Statement

Serendipity, particularly in the biomedical sciences, is known to lead to positive outcomes that can improve the health and well-being of large portions of society (McCay-Peet & Toms, 2018). The role of information science is to connect people with the information they need to accomplish tasks that contribute to the greater societal good. However, providing appropriate support necessitates a complete understanding of the information behavior that one wishes to support. The information science community needs to more fully understand the phenomenon of serendipity in order to effectively facilitate the positive outcomes that serendipity is credited with generating. Information science researcher currently believe that serendipity occurs in a three-step process consisting of an unexpected information encounter followed by time to process the information and resulting in a positive outcome (McCay-Peet & Toms, 2018), however, much of the understanding of serendipity in general, and serendipity in the research environment in particular, is based on example collections (Gries, 2009), which may not provide a balanced, representative overview of serendipity as it is routinely experienced in the course of scientific research. Furthermore, it is important to understand the whole process of serendipity to fully appreciate the impact of research policies, disciplinary traditions and academic reporting practices on this unique type of information behavior. This makes serendipity reporting in scientific research an important topic of study for information scientists.

Theoretical Foundations

The theoretical background for this three-article dissertation is centered on the phenomenon of serendipity as an information behavior. The development of information behavior theory from the initial focus on systems evaluation and user search through the current conceptions of serendipity as a process to be supported is outlined below.

Information and Information Behavior

Information is the data people use to describe, understand and interact with their environment (Rowley, 1998). The concept of information is frequently approached from an interpretivist epistemology, where information is understood as a human construct and the context and experiences of the information user necessarily change the information (Fisher, Erdelez, & McKechnie, 2005, pp. 7–14). Within the interpretivist epistemology, many information behavior theories embrace a constructivist paradigm (Case, 2012, pp. 187–194), where information is seen as describing an external reality, but that "people only know that reality indirectly through their constructions" (Raskin, 2012, p. 119). An excellent example of constructivism in relationship to the concept of information was put forth by Ruben (1992). He describes the concept of information in terms of three "orders." First order information is described as "environmental artifacts and representations; environmental data, stimuli, messages or cues" – those external stimuli to which the individual may, or may not, attend. The second order in Ruben's description involves the internalization of these stimuli, the way in which the individual incorporates the stimuli into their own mental models and appropriates the data for application. The final order that Ruben posits involves the social negotiation of meaning regarding the stimuli, where the individual compares their personal representations with those of others

and modifies their personal understanding based on feedback received. These socially constructed concepts of information have largely influenced the study of information behavior, underpinning the work of many information scientists, including Chatman (social construction of knowledge), Kuhlthau (personal construct theory), Pettigrew (nee Fisher) (discourse analysis), and Dervin (structuration theory) (Case, 2012, pp. 180–194).

Information behavior, likewise, can be viewed through the interpretivist lens as the ways in which people engage with information to construct their understanding of the world. Initial studies of information behavior focused on purposeful behaviors undertaken in response to an information need. Case (2012) notes that information needs, wants or desires result in observable behaviors, but individuals may have difficulty articulating details of their information needs. This purposeful information behavior in response to an information need has been described as information seeking or information search. Yet, not all information is sought after and not all information behavior is purposeful. In creating an inclusive definition of information behavior, Case (2012, p. 5) has described it as an encompassing term that includes "information seeking as well as the totality of other unintentional or passive behaviors (such as glimpsing or encountering information), as well as purposive behaviors that do not involve seeking, such as actively avoiding information."

The multidisciplinary nature of human information behavior has resulted in the identification of many factors that influence user experience. Research in human information behavior emerged from the field of library science in 1948 (Wilson, 2000), and has been informed by computer science studies of user requirements, psychological studies of cognition and information processing, and other fields as disparate as macro-

economics and organizational theory. The influences of these varied fields are encompassed in Wilson's revised general model of information-seeking behavior, where intentionality, context, activating mechanisms, and intervening variables, including psychological, demographic, interpersonal, and environmental factors, shape the user's information processing and use (Fisher et al., 2005, pp. 31–36). While some theories focus on the factors that motivate searches or impact persistence in searching, other theories seek to describe user behaviors as they relate to information outcomes. In these studies, user intentionality in their information behavior and user expectations related to the outcomes realized play key roles in reporting of serendipity.

Intentionality and Expectations in Theories of Information Behavior

Intentionality and expectations are key constructs in identifying an experience as serendipitous. These factors and their interactions can be considered in a 2x2 matrix (see Table 1). Within the field of information behavior, many models focus on the acquisition of information, with most incorporating purposeful or intentional search strategies to arrive at an expected outcome that meets a defined information need.

Table 1. Serendipity Matrix: This table demonstrates the interaction of intentionality of information behavior with expectedness of the outcome.

	Intentional Information Behavior	**Unintentional Information Behavior**
Expected Information Outcome	Intentional search for expected outcomes – Traditional focus of Information Search studies	Finding needed information when not looking for it - Serendipity
Unexpected Information Outcome	Finding unusual or surprising information when looking for something else – Serendipity	Being presented with unusual or surprising information when not looking for anything in particular - Serendipity

However, the process of information acquisition (Erdelez, 2009) can be broken down into intentional acquisition of information (e.g. information seeking) and the unintentional or opportunistic acquisition of information.

Theories of information behavior can be analyzed by the way in which they address the intentionality of the user's information behavior and the way they define the user's expectations regarding the outcome of their information behavior. Using the same 2x2 matrix (see Table 2), we can see that traditional theories of information seeking have defined information behaviors and expected information outcomes. Among these theories is Ellis's Model of Information-Seeking Behavior (1989) which outlines six specific types of intentional activities that users undertake to obtain their expected outcomes. Likewise, the idea of Communities of Practice (Wenger, 2000) describes the culture in which learners acquire information behaviors from practitioners in the field insuring that group members arrive at a shared understanding of community knowledge. Leckie's General Model of the Information Seeking of Professionals (1996) further relates the intentional information behavior of professionals to the work tasks they are undertaking.

Table 2. Models of Information Behavior Related to Serendipity

	Intentional Information Behavior	**Unintentional Information Behavior**
Expected Information Outcome	Ellis's Model of Information-Seeking Behavior (1989) Communities of Practice (Wenger, 2000) General Model of the Information Seeking of Professionals (Leckie et al., 1996)	Information Acquiring-and-Sharing (Rioux, 2005)
Unexpected Information Outcome	**Information Encountering** (Erdelez, 1997) Berrypicking (Bates, 2005) Chang's Browsing (2005)	Ecological Theory of Human Information Behavior (Williamson, 1998) Information Grounds (Fisher, 2005)

While there may be some aspects of these models that allow for the experience of information encounters without intentionality or expectations, the primary focus is on expected outcomes derived from purposeful action.

Initially, theory development related to serendipity arose from the information seeking theories and the need to fit unexpected findings into the models of information search. The key theory related to such unexpected findings is Erdelez's theory of Information Encountering (1997). In the context of information search, Information Encountering describes a mechanism of information acquisition where the seeker, involved in a search, notices and stops to consider information that is intriguing, but unrelated to their immediate purposes. The seeker may choose to examine the serendipitously acquired information more closely or capture the information for later use, or they may just dismiss it. This information encounter interrupts the primary purpose of the user's actions and provides information that is related to some background interest, problem or task. When originally proposed, Erdelez's theory assumed that the user was engaged in intentional information behaviors. It has since become identified as the process undertaken when an individual is presented with a serendipitous stimuli and is frequently applied in both intentional and unintentional search settings. Related models of information behavior include Berrypicking (Bates, 2005) where the focus of an information search evolves with successive information retrieval, and Chang's Browsing (2005) where search is undertaken without a specific goal or information need.

Other theories of information behavior have focused more heavily on the intentionality of the user's behavior when finding this anomalous information. One such theory is the Ecological Theory of Human Information Behavior (Williamson, 1998). In

this theory, Williamson (2005, p. 128) emphasizes that information is "often incidentally acquired rather than purposefully sought." She goes on to describe how users relate to different information sources or systems "plays a significant role in incidental information acquisition" (Williamson, 2005, p. 130). Similarly, Fisher's model of Information Grounds (2005) relates the idea of information flow (Naumer, 2005) – an optimal information experience – to temporally located social interaction. Fisher (2005, pp. 185–186) notes that "As people gather at an information ground, they engage in social interaction, conversing about life, generalities, and specific situations that lead to serendipitous and sometime purposive, formal and informal sharing of information on varied topics." Both Williamson and Fisher provide a framework for information encounters where purposive search is not undertaken and there is no preconceived idea of an information need. Williamson (2005, p. 129) even notes that "some needs are 'unconscious' becoming recognized only when relevant information is discovered".

Finally, individuals may be aware of information needs, but not actively seeking to fill those needs when they encounter the relevant information. The model of Information Acquiring-and-Sharing (Rioux, 2005) is one such theory. In this theory, Rioux introduces the situation of finding information for others. The individual is aware of information needs of others and stores a mental representation of the need. Then, when the individual encounters information related to the need, they capture that information and share it with the individual in need. Rioux (2005, p. 171) also introduces the idea of a "cognitive trigger" where "information of a certain quality (e.g. utility, novelty, interest)" causes the encounterer to associate that information with the stored need.

The theories of information behavior that focus on information acquisition that occurs outside of the intentional search for expected outcomes inform our understanding of serendipity. Whether focused on the unexpectedness of the information encountered or the accidental way in which information is sometimes discovered, the nuances these models reveal bring us closer to understanding serendipity and facilitating its achievement.

Serendipity as an Information Behavior

In synthesizing the existing research on serendipity in information environments, McCay-Peet and Toms (2018, pp. 4–6) have moved us toward a more holistic understanding of serendipity, defining it as an unexpected, anomalous, or inconsistent observation that culminates in a valuable, positive outcome. They have found that serendipity is typically approached as either a quality that describes a person, event or encountered object, or as a process or experience that exhibits one or more serendipitous qualities. According to McCay-Peet and Toms, for an event, outcome or process to be considered serendipitous, an individual, with the knowledge, skills and abilities to capitalize on the information, must experience an information encounter and be willing to invest the time necessary to bring a positive outcome to light. In this series of studies, serendipity will be viewed as a process or experience.

Stages of the Serendipity Experience

McCay-Peet and Toms (2018, pp. 24–26) analyzed five models of serendipity in information behavior (Corneli, Jordanous, Guckelsberger, Pease, & Colton, 2014; Makri & Blandford, 2012; McCay-Peet & Toms, 2015; Rubin, Burkell, & Quan-Haase, 2011; Sun, Sharples, & Makri, 2011). Across the literature, some common features of the

serendipity experience were identified. The first feature is the existence an information encounter. This information, as noted by Rioux (2005, p. 171), is sufficiently unique, useful or otherwise meaningful to prompt the individual to examine the information more closely. Thus, the information encounter serves as a cognitive trigger to the individual and prompts them to actively attend to the information presented. But the serendipitous outcome is not realized just because an individual encounters anomalous information; that individual must also possess the "sagacity" to apply that information in a way that enables them to recognize the value of the information. "Sagacity" is a term associated with serendipity from its inception. Meaning "foresight, discernment, or keen perception" ("sagacity," n.d.), Horace Walpole used the idea of "accidental sagacity" in the letter that coined the term "serendipity" (Foster & Ford, 2003). Regardless of the descriptor employed, it has come to be recognized that serendipity cannot exist apart from a prepared mind. That preparation, along with perseverance and an incubation period, are necessary to achieving a valuable outcome, which may be global or individual in scope. In addition to the common features mentioned above, three of the serendipity models (Makri & Blandford, 2012; McCay-Peet & Toms, 2015; Rubin et al., 2011) include a reflection component following the accomplishment of the outcome in which the individual acknowledges the serendipity of the experience and decides to frame it as such. In this series of studies, the experience of serendipity is generally defined as a process or experience beginning with encountered information that is analyzed and further pursued by an individual with related knowledge, skills and understanding that allows them to realize a valuable outcome. The information encounter that initiates the serendipity experience exhibits qualities of unexpectedness as well as value for the user.

Terminology Describing Serendipitous Information Behavior

Over the course of this study, our understanding of serendipity as an information behavior has grown, and so too has the terminology used to describe it. In Chapter 2, the use of the terminology "ODI" or "opportunistic discovery of information" may be noted. This term is synonymous with information encountering, which is considered to be the initiating stage of serendipity. The reference to the terms used to identify instances of serendipity also changed over time. In the first paper , the synonyms for serendipity were denoted as "ODI terms" or "serendipity search terms." By the completion of the second phase of the study, the term "information encounter" was used to describe the initial phase of serendipity, but the linguistic cue used to identify a reference to serendipity was still being described as a "serendipity synonym" or simply as a "search term." In the final phase of this study, analysis of the grammatical structure of author statements regarding serendipity led to the identification of these linguistic cues as having two aspects. The first is a descriptor term that conveys the unexpected aspect of serendipity, which we have labeled as the "surprise descriptor." It was also determined that the surprise descriptor frequently modifies a noun that portrays the type of information that the researcher found surprising. We have labeled this word the "information term".

Rationale and Contribution

These studies aim to enrich the understanding of serendipity as reported by researchers in the biomedical sciences. Moreover, by using natural language processing techniques in the identification of the corpus of documents to be analyzed, we gain a more balanced and representative understanding of researchers' serendipitous

experiences (Gries, 2009). McCay-Peet and Toms (2018, p. 24) note that "What needs to be augmented in models and frameworks of serendipity are the factors that influence the process or experience. Qualities and characteristics relating to each of the main elements all hold the key to filling out and ultimately helping to combine models of serendipity from various fields of research." This study will contribute to meeting that need.

Research Questions

This series of studies focuses on understanding serendipity as it occurs in the context of biomedical scientific research. The following research questions are addressed.

Research Question 1. How do researchers in biomedical and life sciences use terms related to serendipity in their research reporting?

Research Question 2. What proportion of biomedical research publications indicate serendipitous experiences?

Research Question 3. What syntactic and semantic features distinguish references to serendipitous experiences?

Research Question 4. What features of author reporting are most useful for developing algorithms to automatically compile references to serendipitous experiences?

Methodological Foundations

This three-paper dissertation integrates both quantitative and qualitative analytical techniques to identify patterns of communication regarding serendipity in natural texts. The purpose of this dissertation is to understand the shared experiences of serendipity for biomedical researchers, particularly focusing on how the average researcher experiences and reports serendipity. In achieving this purpose, content analysis has been undertaken

(Krippendorff, 2004). In analyzing the data, an *exploratory concurrent mixed methods approach* (Creswell, 2014, p. 16) has been applied.

Exploratory Concurrent Mixed Methods Approach

These studies are exploratory in that we do not know from the outset which variables will emerge and which will be most important to examine. Because research on serendipity is fairly new and little is known with respect to factors influencing the reporting of serendipitous experiences, an exploratory approach is appropriate.

Concurrent procedures mixed methods procedures, as described by Creswell (2014), allow the researcher to collect both qualitative and quantitative data simultaneously during the study and integrate the information in the interpretation of the overall results. The coded text will be analyzed both qualitatively and quantitatively allowing us to gain a greater appreciation of the impact of different linguistic choices on the aspects of the serendipity experience they reveal. By nesting the quantitative analysis within the larger qualitative analysis, we are able to answer different questions and analyze different units within the study.

Content Analysis

In exploring the language researchers use to describe their serendipitous experiences in research documents, a content analysis will be employed. According to Krippendorff (2004, p. 18), "Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use." In developing the inferences from the texts, several assumptions are applied. First, it is understood that the text has no meaning apart from that created by the reader. There is no message without someone engaging with the texts. It is also understood that

texts have numerous meanings that can be identified by approaching the text from different viewpoints. Furthermore, the meaning conveyed by the texts provide information about events, objects, or ideas that the reader could otherwise not observe. According to Krippendorff (2004), the role of the content analyst is to look outside the physicality of the texts to the conceptions and actions the texts encourage. Content analysis goes beyond computer text analysis because computers do not have the ability to draw inferences outside the data being processed. "When a computer program is used to analyze words, the algorithms that determine the program's operation must embody some kind of theory of how humans read texts, rearticulate texts, or justify actions informed by the reading of texts (Krippendorff, 2004, p. 19)." Creating effective algorithms for identifying instances of serendipity (Erdelez, Marinov, & Allen, 2012), is thus dependent on the development of an effective theory of how humans identify references to serendipity in the literature.

As a method of data analysis, content analysis involves a recognized sequence of procedures. The researcher first determines the units to be analyzed and the sampling procedure to be used. By employing specific sampling strategies, the researcher can be more confident in the validity of the findings. The researcher then determines the recording units, which are described in a coding scheme, and applies those codes to each of the units sampled.

The process of applying codes to the textual data affords a unique option that is not available with other qualitative data collection methods, that of applying both quantitative and qualitative analysis to the data. According to Krippendorff (2004), text is always qualitative to begin with, and a verbal answer to the research question may well

arise from the categorization of textual units. However, the unitization, sampling and coding of the textual data enable quantitative analysis to be performed. Correlations between textual and coded features in the analysis can be used to develop generalizable relational theories. It is, therefore, pragmatic to engage in mixed methods analysis when undertaking content analysis approaches. Throughout this series of studies, content analysis will be employed involving both text and statistical analysis of full-text research articles mining for synonyms for serendipity.

Content analysis is an accepted method in serendipity research. Due to the transient nature of serendipitous experience, data collection can prove difficult. One option for studying these transient experiences is to use existing documents as a source for studying instances of serendipity. Rubin et al (2011) applied content analysis to blog posts to further their understanding of serendipity. Likewise, Campanario (1996) employed content analysis to reveal serendipity in researchers' commentaries on highly cited papers.

The Dissertation Format

This dissertation features three phases of investigation aimed at describing serendipity as it is reported in the biomedical research literature. Each phase is presented as a separate article, with its own abstract, introduction, methods, findings and conclusions. The first article, presented in Chapter 2, entitled "Looking for Opportunistic Discovery of Information in Recent Biomedical Research – A Content Analysis" was published in *Proceedings of the American Society for Information Science and Technology.* That article presents a taxonomy of interpretations that were conveyed by the authors of biomedical research articles when using phrases that were considered

synonymous with serendipity. The manuscript presented in Chapter 3, entitled "Distraction to Illumination - Mining Biomedical Publications for Serendipity in Research" has also been accepted for publication in *Proceedings of the American Society for Information Science and Technology.* That article describes a refinement in identifying terms for the recognition of reports of serendipity, presents initial logistic regression results, and identifies four variables that demonstrated statistical reliability in predicting a reference to serendipitous events. Finally, Chapter 4 presents a manuscript entitled "Characterizing Serendipity in Biomedical Research Literature: Identification of Syntactic and Linguistic Features" that has been prepared for submission to *Journal of the Association for Information Science and Technology.* That article identifies three syntactic features that are statistically reliable for predicting serendipity, extrapolates findings to estimate a population of 225,000 articles mentioning a serendipitous event indexed by PubMed Central, and presents further refinements of the taxonomy developed in Ch. 2. Additionally, this manuscript provides an example of using natural language processing tools in the discovery of semantic and syntactic indicators of a concept.

A three-article dissertation format was selected because the iterative nature the *exploratory concurrent mixed methods approach* (Creswell, 2014, p. 16) selected for the data analysis lends itself to conducting the study in three phases. The three-article dissertation structure also allows me to gain greater experience with the academic publication process, a key factor in developing a successful academic career. Additionally, as a non-traditional PhD student, the publication of these articles will help me establish a research agenda that I can build upon throughout my career.

References

Bates, M. J. (2005). Berrypicking. In *Theories of Information Behavior* (pp. 69–74). Medford, N.J: American Society for Information Science and Technology.

Campanario, J. (1996). UsingCitation Classics to study the incidence of serendipity in scientific discovery. *Scientometrics*, *37*, 3–24.

Case, D. O. (2012). *Looking for Information : a survey of research on information seeking, needs and behavior*. Bingley, UK: Emerald Group Pub. ;

Chang, S.-J. (2005). Chang's Browsing. In *Theories of Information Behavior* (pp. 69–74). Medford, N.J: American Society for Information Science and Technology.

Corneli, J., Jordanous, A., Guckelsberger, C., Pease, A., & Colton, S. (2014). Modelling serendipity in a computational context. *ArXiv:1411.0440 [Cs]*. Retrieved from http://arxiv.org/abs/1411.0440

Creswell, J. W. (2014). *Research design: qualitative, quantitative, and mixed methods approaches* (4th ed). Thousand Oaks: SAGE Publications.

Ellis, D. (1989). A behavioural model for information retrieval system design. *Journal of Information Science*, *15*, 237–247.

Erdelez, S. (1997). Information encountering: a conceptual framework for accidental information discovery (pp. 412–421). Presented at the ISIC '96: Proceedings of an international conference on Information seeking in context, Taylor Graham Publishing.

Erdelez, S. (2004). Investigation of information encountering in the controlled research environment. *Information Processing & Management*, *40*, 1013–1025.

Erdelez, S. (2009). Information Encountering. In *Theories of Information Behavior* (pp.

179–184). Medford, N.J.: Information Today Inc.

Erdelez, S., Marinov, M., & Allen, C. (2012). Knowledge Discovery System for Research Hypothesis Generation from Serendipitous Findings: A Feasibility Study. In *Proceedings - AMIA 2012 Annual Symposium*. Chicago: American Medical Informatics Association.

Fisher, K. E. (2005). Information Grounds. In *Theories of Information Behavior* (pp. 185–190). Medford, N.J: American Society for Information Science and Technology.

Fisher, K. E., Erdelez, S., & McKechnie, L. (2005). *Theories of information behavior*. Medford, N.J.: Published for the American Society for Information Science and Technology by Information Today.

Foster, A., & Ford, N. (2003). Serendipity and information seeking: an empirical study. *Journal of Documentation*, *59*, 321–340.

Gries, S. T. (2009). *Quantitative corpus linguistics with R : a practical introduction*. New York, NY : Routledge, 2009. Retrieved from https://www.safaribooksonline.com/library/view/quantitative-corpus-linguistics/9781135895594/Part01.html

Krippendorff, K. (2004). *Content analysis : an introduction to its methodology*. Thousand Oaks, Calif.: Sage.

Leckie, G. J., Pettigrew, K. E., & Sylvain, C. (1996). Modeling the Information Seeking of Professionals: A General Model Derived from Research on Engineers, Health Care Professionals, and Lawyers. *The Library Quarterly: Information, Community, Policy*, *66*, 161–193.

Makri, S., & Blandford, A. (2012). Coming across information serendipitously – Part 1:

A process model. *Journal of Documentation*, *68*, 684–705.

Marton, F. (1986). Phenomenography—A Research Approach to Investigating Different Understandings of Reality. *Journal of Thought*, *21*, 28–49.

McCay-Peet, L., & Toms, E. (2018). *Researching serendipity in digital information environments*. Retrieved from http://dx.doi.org/10.2200/S00790ED1V01Y201707ICR059

McCay-Peet, L., & Toms, E. G. (2015). Investigating serendipity: How it unfolds and what may influence it. *Journal of the Association for Information Science and Technology*, *66*, 1463–1476.

Naumer, C. (2005). Flow Theory. In *Theories of Information Behavior* (pp. 153–157). Medford, N.J: American Society for Information Science and Technology.

Raskin, J. D. (2012). Evolutionary constructivism and humanistic psychology. *Journal of Theoretical and Philosophical Psychology*, *32*, 119–133.

Rioux, K. (2005). Information Acquiring-and-Sharing. In *Theories of Information Behavior* (pp. 169–173). Medford, N.J: American Society for Information Science and Technology.

Rowley, J. (1998). What is information? *Information Services & Use*, *18*(4), 243.

Ruben, B. D. (1992). The communication-information relationship in system-theoretic perspective. *Journal of the American Society for Information Science*, *43*, 15–27.

Rubin, V. L., Burkell, J., & Quan-Haase, A. (2011). Facets of serendipity in everyday chance encounters: a grounded theory approach to blog analysis. *Information Research*, *16*, 27–27.

sagacity. (n.d.). *Dictionary.com Unabridged*. Retrieved from

http://www.dictionary.com/browse/sagacity

Sun, X., Sharples, S., & Makri, S. (2011). A user-centred mobile diary study approach to understanding serendipity in information research. *INFORMATION RESEARCH-AN INTERNATIONAL ELECTRONIC JOURNAL*, *16*. Retrieved from http://informationr.net/ir/16-3/paper492.html

Wenger, E. (2000). Communities of Practice and Social Learning Systems. *Organization*, *7*, 225–246.

Williamson, K. (1998). Discovered by chance: The role of incidental information acquisition in an ecological model of information use. *Library & Information Science Research*, *20*, 23–40.

Williamson, K. (2005). Ecological Theory of Human Information Behavior. In *Theories of Information Behavior* (pp. 128–137). Medford, N.J: American Society for Information Science and Technology.

Wilson, T. D. (2000). Human information behavior. *Informing Science*, *3*, 49–55.

Chapter 2 Looking for Opportunistic Discovery of Information in Recent Biomedical Research – A Content Analysis

This chapter was originally published as:

Allen, C. M., Erdelez, S., & Marinov, M. (2013). Looking for opportunistic discovery of information in recent biomedical research – a content analysis. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–11. https://doi.org/10.1002/meet.14505001082

Abstract

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'

— Isaac Asimov

From the discovery of penicillin and x-rays to the development of many of today's chemotherapy agents, serendipitous findings tangential to the researcher's intended purpose, those "*That's funny…*" moments, have greatly impacted the health and well-being of society. As an information behavior, these unexpected findings are an example of the Opportunistic Discovery of Information (ODI). ODI has been described in many contexts, from information behavior in virtual worlds to the impact of information encountering on health behaviors. Yet, little is known about instances of ODI within the context of scientific research. This study uses content analysis to reveal reported instances of ODI in recently published biomedical literature. Our findings propose a taxonomy of term use indicating the presence of serendipity in the research process and

reveal the relationship between the authors' word choice for serendipity and specific types of ODI experiences.

Introduction

Imagine for a moment, a large research university where a researcher in exercise physiology is studying the impact of an intervention on subjects' lipid levels. In reviewing his subjects' lab results, he notes that several subjects are demonstrating increased creatinine levels, but the intervention was not supposed to impact the kidneys. Across campus, a researcher studying the connection between cadmium and endometrial cancer observes that several of her subjects have increased white cell counts unrelated to the cancer incidence. Meanwhile, a radiologist on the health sciences campus has noted an increase in the number of cystic liver lesions occurring incidentally on high resolution chest CTs. Upon further investigation, she notes that all of the patients are being seen in the same clinic. You might wonder how often researchers run into unexpected or surprising findings tangential to their research in the normal course of research performance.

Background

The idea of accidental discoveries made while in search of other information is a growing research focus for information scientists. The field of human information behavior (HIB) addresses this idea through the functional Model of Information Encountering (Erdelez, 2009). In the context of information search, Erdelez's Information Encountering (1999), describes a mechanism of information acquisition where the seeker, involved in a search, notices and stops to consider information that is intriguing, but unrelated to their immediate purposes. The seeker may choose to examine

the serendipitously acquired information more closely or capture the information for later use, or they may just dismiss it.

The concept of Opportunistic Discovery of Information is broader than Information Encountering. It also includes other types of what is commonly referred to as serendipity that are yet not fully explored in HIB literature. A review of the literature regarding Opportunistic Discovery of Information reveals studies in the realms such as geospatial imaging (Smith, 2011), information literacy (Erdelez, Basic, & Levitov, 2011), web searching (Miwa et al., 2011), online consumerism (Wang et al., 2011), and news reading (Yadamsuren & Erdelez, 2010). Most current studies of information encountering have been focused on the discovery of text-based information. Yet, if we define information as "any difference you perceive, in your environment or within yourself" (Case, 2012, p. 4), it seems reasonable that such accidental discoveries can occur during engagement in virtually any life activity. One activity that focuses on the search for information is scientific research. With the explosion in the volume of data that can be captured, processed, and analyzed, the possibility of encountering unexpected information during research performance is growing. Indeed, several notable medical advancements have resulted from recognized instances of ODI during scientific studies with other foci (Barnett, 2011; Hargrave-Thomas, Yu, & Reynisson, 2012; Lee, 2011; Ligon, 2004; Mayor, 2010; Mould, 1995; Rubanyi, 2011; Young, Ashdown, Arnold, & Subramonian, 2008). However, little is known about the experience of ODI across the research community.

A major difficulty in the study of the Opportunistic Discovery of Information is the transient nature of the experience. People do not plan to find information

unexpectedly, nor is it easy to develop an experimental environment that consistently fosters the ODI experience (Erdelez, 2004). The unpredictable aspect of these ODI experiences makes direct observation difficult, if not impossible, necessitating the utilization of indirect research approaches. The majority of ODI studies have employed interview or survey tools to investigate the participants' recollections or impressions of ODI events with significant success, but it would complement the understanding of ODI to approach the phenomenon from additional methodological designs. Content analysis is particularly useful for questions that are "believed to be answerable by examination of the body of texts," and that "concern currently inaccessible phenomena (Krippendorff, 2004, pp. 32–33)." Content analysis, therefore, may prove to be a useful methodology in revealing instances of Opportunistic Discovery of Information. In content analysis, documents, such as journal articles, can be analyzed for both their manifest content (word use or count) and their latent content (themes and meanings). We believe that the current research literature holds latent references to these ODI experiences and can be systematically analyzed to reveal traces of this human information behavior.

Rationale and Contribution

The purpose of this study is to seek evidence of Opportunistic Discovery of Information in the context of performing biomedical research. This study focuses on the following questions:

Research Questions

1. How are terms related to serendipity used within biomedical research literature?

2. Given a particular synonym for serendipity, what is the probability that the author is referring to an experience of Opportunistic Discovery of Information?
3. How prevalent is the Opportunistic Discovery of Information in recent research publications?

Methods

To answer the posed questions, a content analysis was undertaken.

Sampling Units

In this study, the units sampled are journal articles known to contain synonyms for serendipity. However, in content analysis, the units sampled do not equate to the units counted. The context units and the recording units are defined below.

Study Population

The analysis sample was drawn from full text journal articles indexed in PubMed Central. PubMed Central was selected for its extensive collection of over two million full-text, primary research reports. PubMed Central is the free archive of biomedical and life sciences journal literature compiled by the U.S. National Institutes of Health's National Library of Medicine. The ability to search the full-text of the articles without the need for continual access to journal subscription services, as well as the extensive range of articles indexed made PubMed Central a logical starting point for this investigation.

It was important to know if the idea of the Opportunistic Discovery of Information was currently indexed by the existing PMC user tools. In addition to full text searching, PubMed Central offers users the ability to search by Medical Subject Headings (MeSH terms). MeSH headings are frequently used to narrow literature searches to the user's key

ideas. However, the MeSH category designation requires the documents to be classified with these terms by an expert examiner and assigned MeSH terms are limited to 10 to 12 terms per document, so minor ideas contained within the articles are not indexed by the MeSH terms. The current PubMed Central ontology does include the MeSH term "incidental findings," which includes the identified synonyms for serendipity: "incidental finding(s)", "finding(s), incidental", "incidental discovery(ies)", and "discovery(ies), incidental". A search of PubMed Central by usage of this MeSH heading returned only 303 articles. So, while PubMed Central has a MeSH term that is mapped to the idea of serendipity, use of the MeSH heading fails to adequately return instances of Opportunistic Discovery of Information, revealing the necessity of full text search for the related terms.

Context Units

In order to determine which articles within the PubMed Central database might contain information related to the Opportunistic Discovery of Information, we had to identify the context units, i.e., the textual matter upon which the analysis will focus. For this study, the context units are the synonyms related to the idea of serendipity. Synonyms for serendipity were derived through a brainstorming process undertaken by members of the research team. After the initial list of synonyms was created it was validated by information behavior researchers outside the project. Terms identified through this process include: accidental discovery(ies), accidental finding(s), chance discovery(ies), chance finding(s), fortuitous discovery(ies), fortuitous finding(s), incidental discovery(ies), incidental finding(s), serendipitous discovery(ies), serendipitous finding(s), unanticipated discovery(ies), unanticipated finding(s), unexpected discovery(ies), and unexpected finding(s).

A full text search was performed for each of the identified terms with results filtered to include only research and review articles. The searches returned a total of 23,571 articles with frequencies as described in Table 1. Unexpected finding(s) was by far the most commonly used term related to serendipity. However, without further analysis of the semantic meaning of the term used, it was unknown how useful the term would be for returning incidents of Opportunistic Discovery of Information.

Sampling

Because we believe that presence of the identified context units will provide us with the greatest probability of finding evidence of ODI, we did not randomly sample the entire population of articles within the PubMed Central database; we, instead, utilized purposive sampling to target those articles most closely related to ODI. As that population of 23,571 exceeded what could reasonably be analyzed, a sample of 436 articles from that relevant pool was selected based on a table of recommended sample sizes for populations with finite sizes (Patten, 2005, p. 179). This sample provides 95% confidence with a precision (half-width of the interval) of 0.1. In order to most accurately compare the characteristics across the subgroups of search terms, a stratified sampling method was also employed. Equal numbers of articles were drawn from the pool of articles returned for each search term. Articles were drawn from the most recent publications for each term. We felt that drawing from the most recent publications would provide us with information concerning the current status of ODI experiences. Each article was analyzed with respect to the use of the term for which that article was selected. While the articles could contain more than one search term, only the context unit for which the article was selected was analyzed. No duplicate articles were drawn, so replacement was not necessary.

Table 1. Incidence and search details for synonyms of serendipity.

Search Term (Context Unit)	**# of Articles**	**Search Statement**
Accidental discovery(ies)	330	"accidental discovery"[Text Word] OR "accidental discoveries"[Text Word] AND "research and review articles"[filter]
Accidental finding(s)	341	"accidental finding"[Text Word] OR "accidental findings"[Text Word] AND "research and review articles"[filter]
Chance discovery(ies)	177	"chance discovery"[Text Word] OR "chance discoveries"[Text Word] AND "research and review articles"[filter]
Chance finding(s)	2242	"chance finding"[Text Word] OR "chance findings"[Text Word] AND "research and review articles"[filter]
Fortuitous discovery(ies)	145	"fortuitous discovery"[Text Word] OR " fortuitous discoveries"[Text Word] AND "research and review articles"[filter]
Fortuitous finding(s)	121	"fortuitous finding"[Text Word] OR "fortuitous findings"[Text Word] AND "research and review articles"[filter]
Incidental discovery(ies)	206	"incidental discovery"[Text Word] OR "incidental discoveries"[Text Word] AND "research and review articles"[filter]
Incidental finding(s)	4830	"incidental finding"[Text Word] OR "incidental findings"[Text Word] AND "research and review articles"[filter]
Serendipitous	2171	"serendipitous "[Text Word] OR "serendipitous "[Text Word] AND "research and review articles"[filter]
Serendipitous discovery(ies)	391	" serendipitous discovery"[Text Word] OR " serendipitous discoveries"[Text Word] AND "research and review articles"[filter]
Serendipitous finding(s)	245	" serendipitous finding"[Text Word] OR " serendipitous findings"[Text Word] AND "research and review articles"[filter]
Unanticipated discovery(ies)	31	" unanticipated discovery"[Text Word] OR " unanticipated l discoveries"[Text Word] AND "research and review articles"[filter]
Unanticipated finding(s)	489	" unanticipated finding"[Text Word] OR " unanticipated findings"[Text Word] AND "research and review articles"[filter]
Unexpected discovery(ies)	459	" unexpected discovery"[Text Word] OR " unexpected discoveries"[Text Word] AND "research and review articles"[filter]
Unexpected finding(s)	11,384	" unexpected finding"[Text Word] OR “unexpected findings"[Text Word] AND "research and review articles"[filter]
Total	**23,571**	

Coding Units and Coding Scheme Development

The purpose of this study was determine how authors of biomedical research use terms related to serendipity in hopes that some of the meanings conveyed by these terms would provide indication of ODI experiences. It was not the purpose of this study to determine if the findings described as unusual or surprising were truly unique in their respective fields. We view the researchers who published the articles as professionals capable of recognizing surprising findings in their fields and sought only to analyze the way in which they used the terms which could be relevant to ODI. Therefore, the researchers in this study have background in information behavior, but do not purport to be experts in all the fields represented in this study. The procedure for identifying the recording units and developing the coding scheme involved the following steps:

1. An initial set of recording units was developed based on the researchers' general knowledge and conceptual understanding of the opportunistic discovery of information.
2. These categories were critically analyzed, discussed, and refined by the entire research group.
3. This a priori set of meanings was then applied by the coders to a sample of the articles. This application resulted in the discovery of several additional categories of term use and led to further refinement of the category definitions, until it was felt that the identified categories were both exhaustive and mutually exclusive.

This process resulted in the coding scheme presented in Table 2.

Table 2. Recording units used for coding with explanations.

Categories	Coding Units	Description
Relevant to ODI	Inspiration	ODI forms the basis of the research design; typically a study to follow up a serendipitous finding of a previous study.
	Research Focus	ODI constitutes the major research findings of the study; report is focused on the serendipitous finding, rather than the other research goals of the study.
	Mentioned Findings	ODI findings resulted from the study, but were not the major research outcomes; report primarily focuses on the initial research questions, but mentions some instances of serendipity or unexpectedness.
	Systematic Reviews	Meta-analyses of ODI contributions to a particular field
Irrelevant to ODI	Historical Reference	Serendipity term is mentioned in reference to another study as part of the background; usually located in the literature review or background section.
	Statistical Reference	Serendipity term was used to refer to significance in statistical testing
	Non-Research Focus	Serendipity term is the focus of the article, but the article does not disseminate new research findings, i.e. a historical review or letter to the editor.
	Further Study	Serendipity term is included as an area for further study without elaboration on the specific findings
	Conveyance of Insignificance	Serendipity term is used as an adjective to convey insignificance or inconsequentiality, usually in reference to the medical identification of a disease, but occasionally in other contexts
	Reference Title	ODI term was found in titles from the reference list

Coding

After the schema development and practice described above, two coder/researchers independently applied the coding scheme to all of the remaining articles in the sample.

A subset of 100 articles was analyzed for inter-coder reliability using Krippendorff's Alpha coefficient. Krippendorff's alpha has a reputation for being a highly accurate method of determining inter-coder reliability, and the use of an open source SPSS Macro (www.afhayes.com) streamlined the calculation process. The results indicate a significant level of agreement between the coders.

KAPLHA = .9001

All disagreements were resolved by independently recoding the units in disagreement, which resolved the majority of disagreements. All remaining disagreements were discussed and coded by consensus.

Findings

The first objective of this study was to determine how biomedical authors use synonyms related to serendipity in their writing, and which of these usages seemed to indicate the presence of an ODI experience. The coding scheme presented in Table 2 was used to classify the term use as ODI relevant or ODI irrelevant, as well as to specify the way in which the term was used.

ODI Relevant Term Use

Of the ten distinct usages of the serendipity search terms identified, four categories of usage were determined relevant to ODI, as the reference provided additional facts or knowledge useful for understanding the phenomenon. The categories of usage deemed relevant were Inspiration, Research Focus, Systematic Reviews and Mentioned Findings.

Inspiration

The Inspiration category is comprised of articles where the serendipity described forms the basis of the research design or seeks to further describe previously recorded ODI phenomena. The inspiration reference related to

Inspiration

This work *was motivated by* the fortuitous discovery of mtDNA length heteroplasmy in crickets while I was learning how to use mtDNA to study the Gryllus hybrid zone in Rick Harrisons' lab.

Rand (2011)

As a continuation of the previous findings in human fetuses, accidental finding of an accessory vascular component in the posterior part of CAC of human adult cadavers *inspired the authors to* present and compare its posterior part configuration.

Vasović et al. (2010)

Figure 2.1 Inspiration

ODI term use can be seen in the following quotes retrieved for this study.

The search terms “accidental finding” and “fortuitous finding” returned the most instances of serendipity term use as inspiration.

Research Focus

When the entire article is devoted to reporting a serendipitous finding encountered during the conduct of research, diagnostics or medical therapeutics for another purpose, we considered the term usage to be research focus. These articles were more likely to use multiple synonyms for serendipity and/or use the same serendipity terminology more than once in the article. These articles frequently take the form of case studies.

Research Focus

In this report, we describe a fortuitous discovery of unsuspected lung adenocarcinoma in surgical resection performed for aspergilloma of the right upper lobe.

Smahi, et al (2011)

The present study describes the unanticipated finding that nuclear budding/micronucleation is coupled with cytoplasmic membrane blebbing.

Utani, Okamoto, & Shimizu (2011)

Figure 2.2. Research Focus

Research Focus use of serendipity terms were most frequently returned with searches on “unexpected discovery” and “unanticipated discovery”.

Systematic Reviews

A small group of articles focused on systematic reviews of the incidence of serendipity as it related to a particular field. This category of terminology use was not considered relevant for

Systematic Reviews

Note that the *full range of literature* was not captured by our method, since we chose to use only one search string rather than running a comprehensive search using all possible relevant strings such as “accidental findings” and “unexpected findings” and related Medical Subject Heading (MeSH) terms.

Illes & Chin (2008)

We also performed a *literature review* on PubMed database, limited to the English language, using the following terms: "gastrointestinal stromal tumor," "adnexal mass," and "incidental finding."

Muñoz et al (2012)

Figure 2.3. Systematic Reviews

the development of further research, but was deemed useful in informing the research design of future ODI studies.

Examples of serendipity terminology use from their articles are as follows:

Systematic reviews were most frequently returned in searches using “accidental findings” and “incidental finding”.

Mentioned Findings

The largest incidence of relevant ODI term use was found as mentioned findings. Mentioned findings were determined to be instances where the ODI findings described resulted from the study, but were not the major research outcomes. While the context and tenor of the term use is very similar to the statements that indicated a research focus on the ODI phenomenon, the location of the statements within the paper played a key role in determining whether the ODI incident was the focus of the paper or a mentioned finding. The statements related to a research focus were primarily located early in the paper, in abstracts, introductions and statements of purpose. Mentioned findings, however, were not seen until the results, discussion, or conclusion sections.

Mentioned findings were most frequently returned with searches using “chance finding,” “unanticipated finding” and “unexpected finding”.

Mentioned Findings

This generalization to the dominant eye is perhaps our most unanticipated finding. It is also of considerable clinical relevance, since most strabismic and many anisometropic amblyopes rely mainly on the fellow eye in everyday living, as vision in the amblyopic eye is completely or partially suppressed.

Suttle et al. (2011)

Average birth weight was 296 g higher (95% CI, 109–482 g) in infants born during the cold season (after harvest) than in other infants; this unanticipated finding may reflect the role of maternal nutrition on birth weight in an impoverished region.

Thompson et al. (2011)

Figure 2.4. Mentioned Findings

ODI Irrelevant Term Use

Six categories of usage were determined irrelevant to further research as the ODI reference did not provide additional understanding. Categories of usage deemed irrelevant include Reference Title, Statistical Reference, Conveyance of Insignificance, Historical, Non-Research Focus, and Further Study.

Reference Title

The least relevant instances of serendipity terminology returned were those where the term only appeared in the title on the list of references for the article. Reference titles were most frequently returned in searches on "accidental findings".

Statistical Reference

Many times the serendipity terminology was used to refer to significance in statistical testing. For example, the usage in the Buechel, et al (2011) article was determined to be a statistical reference: "To determine a single Type I error cutoff (α level) for both studies, we constructed fold-enrichment graph depicting the relative increase over chance discovery that real data comparisons show." Similarly, the work of Badgaiyan & Wack (2011) used serendipity terminology in a statistical manner: "To ensure that this measurement reflected endogenously released dopamine and it was not a chance finding, we measured additional receptor kinetic parameters using the E-SRTM."

Articles using ODI terminology in reference to statistical analysis were most frequently returned from "chance findings" searches.

Conveyance of Insignificance

At times the serendipity term is used as an adjective to convey insignificance or inconsequentiality, usually in reference to the medical identification of a disease, but

occasionally in other contexts. “These anomalies can present early in life, or may be just incidental findings (Sundarakumar, 2011).” “Intraspecific competition may be involved, differences in hunting and/or collecting skills and strategies, acquired through learning or chance discovery, could be the reason, and there could even be an outwardly not visible physiological basis for such kinds of behavior (Meyer-Rochow, 2009).”

Insignificance was most frequently returned from searches involving the term “incidental”.

Historical

Historical usage was defined as instances where the serendipity term is mentioned in reference to another study as part of the background. For example, the instance from Barreiro, Martin & Garcia-Estrada’s (2012) work on proteomics was determined to be historical usage. “The history of these compounds started up in 1928 after Sir Alexander Fleming's accidental discovery of the antimicrobial activity generated by a fungus culture contaminating a Petri dish cultured with Staphylococcus sp.” Historical references to serendipity were revealed with searches involving “discovery”, with “accidental discovery” and “serendipitous discovery” returning the most historical articles, followed closely by “fortuitous discovery”.

Non-Research Focus

Occasionally, the article in which the terminology was used was found to have a non-research focus, where the ODI incident forms the focus of the article, but the article does not disseminate new research findings, i.e. a historical review or letter to the editor. For example, in a letter to the editor in response to criticism of their published findings, Hough & Hennekam (2009) made the following statement: “The chance discovery of

concomitant non-symptomatic esophageal papillomatosis was a major additional clue for this." Non-research focus articles were returned exclusively with the search term "chance discovery."

Further Study

Passing mention of serendipity findings as an area for further study without elaboration was determined to provide insufficient information for future research. An example of such passing mention occurs in the Tluczek, et al (2010) article as follows: "The group differences in rates of breast feeding were serendipitous findings that point to the need for additional inquiry about mothers' decision-making process about feeding following a neonatal diagnosis." Although initial review determined further study usages to be irrelevant to the development of ODI driven research hypotheses, further attention should be paid to this category to evaluate that determination. Articles classified as further study were returned most frequently in searches using the term "serendipitous findings".

Relationship of Word Choice to ODI

After determining the classification of authors' term usage, our second objective was to look for correlations between term choice and indications of ODI experiences. Ten distinct usages of the serendipity search terms were identified. Each category was analyzed to identify the search terms most frequently contributing to articles in the category. Chi-Square Tests were performed to determine the relationship between the author's word choice related to serendipity and the relevance of that word to experiences of the opportunistic discovery of information. The first test examined the relationship between ODI relevance and the entire context unit.

For the purposes of this test, the null hypothesis was:

H_O: Author word choice and relevance to ODI are independent.

The relation between these variables was significant, X^2 (28, N = 413) = 133, $\rho <$ .05. Terms associated with relevance to ODI and the percentage of relationship is reported in Table 3. Table 3 presents two different relationships to ODI – the first relationship presented (center column) shows the percentage of all ODI-relevant articles that were derived from each term; the second relationship presented (right column) is the percentage of articles within each term strata that had usages relevant to ODI.

Table 3. Relationship of Word Choice to ODI Relevance

Serendipity Term (Context Unit)	**% of Total ODI Contributed by Term**	**% of Term Instances with Relevant Use**
Accidental discovery(ies)	1.6%	25%
Accidental finding(s)	12.5%	55%
Chance discovery(ies)	2.8%	25%
Chance finding(s)	9.1%	40%
Fortuitous discovery(ies)	3.4%	30%
Fortuitous finding(s)	4.5%	24%
Incidental discovery(ies)	3.4%	30%
Incidental finding(s)	6.2%	28%
Serendipitous	3.4%	30%
Serendipitous discovery(ies)	1.6%	10%
Serendipitous finding(s)	3.3%	33%
Unanticipated discovery(ies)	6.8%	65%
Unanticipated finding(s)	15.3%	83%
Unexpected discovery(ies)	6.2%	60%
Unexpected finding(s)	15.3%	83%

Frequencies by Synonym within ODI- Relevant Articles

In the center column, we see that all of the synonyms for serendipity contribute to the pool of ODI relevant articles. We also see that "accidental finding(s)," "unanticipated finding(s)," and "unexpected finding(s)" make up over 42% of the total number of ODI relevant articles identified in our sample.

ODI-Relevance Frequencies within Each Synonym Sample

The right column of Table 3 looks at the proportion of articles sampled within each term that were classified as ODI-relevant. In this column, we again see that some ODI relevance exists within each of the synonyms for serendipity. We also see that some terms, such as "fortuitous finding(s)," "serendipitous discovery(ies)," and "accidental discovery(ies)" appear to be relatively unfruitful when looking for instances of ODI, while terms such as "unanticipated finding(s)," and "unexpected finding(s)," provide a high rate of ODI relevance.

As the analysis of the terms we actually searched, was performed, we wondered about the relationship of the adjective in each search phrase to ODI relevance. So, a second round of analysis was undertaken to look at the relationship between ODI relevance and the descriptive modifiers from each of the context units. For this test, the search terms were grouped by the adjective in the search term. For example, "accidental discovery," "accidental finding," and "accidental findings" were grouped together under "accidental." For this test, the null hypothesis was:

H_O: Descriptive modifier use and relevance to ODI are independent.

The relation between these variables was significant, X^2 (12, N = 413) = 103, ρ < .000. Terms associated with relevance to ODI and the percentage of relationship is reported in Table 4.

Table 4. Relationship of Modifier to ODI Relevance

Descriptive Modifiers Associated with Context Units	% of Total ODI Contributed by Term	% of Terms Using Modifier with ODI Relevance
Accidental	14.8%	58%
Chance	11.9%	48%
Fortuitous	8.0%	49%
Incidental	9.7%	47%
Serendipitous	11.9%	49%
Unanticipated	22.2%	77%
Unexpected	21.6%	75%

Frequencies by Modifier within ODI- Relevant Articles

The center column of Table 4 demonstrates the distribution of modifiers within the pool of relevant use. In this column, we see that terms using "unanticipated" and "unexpected" make up over 43% of the total number of ODI relevant articles identified in our sample.

ODI-Relevance Frequencies within Each Modifier Sample

The right-hand column of Table 4 looks at the proportion of articles sampled within each modifier that were classified as ODI-relevant. In this column, we see that searches including the modifiers "unanticipated" or "unexpected" return an instance of ODI over 75% of the time.

Prevalence of ODI in the Biomedical Literature

The third objective of this study was to attempt to determine the prevalence of ODI references within recent biomedical literature indexed by PubMed Central.

Frequencies within the Population

By multiplying the percentage of ODI-relevant articles in the sample by the total number of articles returned for each search term, we estimated the number of total number of ODI-relevant articles indexed by PubMed Central. Of the 23,571 articles

returned by the initial search strings, it can be estimated from the sample that 10,900 articles contain ODI-relevant information.

Figure 1 illustrates the estimated frequency of articles describing relevant ODI experiences in the PubMed Central population. Note the break in the illustration for unexpected finding(s). This search string alone can account for more than 86% of the ODI-relevant articles. The least useful search terms were "fortuitous findings" (returning only one relevant article), "serendipitous findings" (returning an estimated 14 articles) and "unanticipated discovery" (returning an estimated 20 articles).

Discussion

The opportunistic discovery of information in contexts outside of the realm of text-based resources can be challenging to define. While some might argue that the identification of unexpected and surprising findings are the natural outcome of carefully planned and conducted research, others would contend that, while less than serendipitous, careful planning and analysis do not negate the opportunistic nature of these findings. A correlation can be made to a carefully planned and conducted literature search. An investigator may carefully select the database and construct search terms to produce literature on a chosen topic, say visual attention, and still return information that is pertinent to a different information need in their lives, for example, amblyopia, a

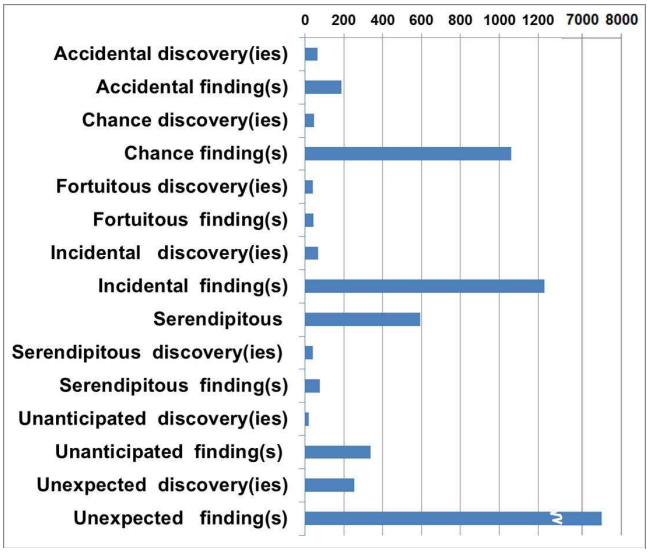

Figure 2.5. Estimated Frequency of ODI Relevant Articles

childhood vision disorder. So, the ODI experience consists of 2 basic components: the identification of the presence of information that is outside of what was originally expected and not dismissed, and the connection of that information to a different problem than the one that originally produced it. While we have proposed some classes that appear to us to be relevant to ODI, some of the instances may only contain one of the 2 components of a complete ODI experience.

Our findings clearly demonstrate that experiences relevant to the Opportunistic Discovery of Information are commonly reported in recent biomedical literature. This is important to the Human Information Behavior literature in that it confirms that ODI experiences occur in contexts beyond text-based sources. Our study applied content analysis methodology to a sample of biomedical articles to reveal evidence of ODI. The reports of ODI in the research papers provide an authentic record of researchers' experiences. While these records do not typically provide much information about the events surrounding the serendipitous discovery, they do provide evidence of the phenomenon and a starting place for additional interviews, surveys and other methods to delve deeper into the ODI experiences of researchers.

Additionally, we found that full text searching for concepts related to serendipity reveals more instances of ODI than using the related MeSH terms. This is important for informing future searches for ODI in research literature. It may also prove helpful in informing modifications to the way the serendipity MeSH term is populated.

Concerning the authors' intent when using terms related to serendipity, we found that authors use terms related to serendipity to convey ten unique meanings. Of these meanings, four are related to the experience of ODI. This finding is helpful in

understanding how ODI is addressed in the biomedical literature and which terms are most helpful in locating incidents of ODI. Overall, the largest percentage of ODI-related articles was returned by the term "unexpected finding(s)."

Our findings have theoretical and practical implications. First, they inform our understanding of how biomedical researchers serendipitously acquire information in a non-text-based context and contribute to expanded application of human information behavior theories in these environments. Additionally, the findings are also useful in informing the design of a data mining system that would facilitate novel research hypothesis generation in biomedical and other areas of research (Erdelez, Marinov, & Allen, 2012). Such a system could be developed to connect unexpected, orphaned findings from one research study with researchers who possess knowledge and skills to act upon those findings.

Limitations of the Study

While we estimate, based on this sample, that nearly half of the articles containing synonyms for serendipity will describe some instance of ODI, this percentage may be an aberration of the timeframe analyzed and may not be applicable to all PubMed Central articles using serendipity terms. However, the use of phrases rather than individual terms as synonyms for serendipity may have resulted in a much lower estimate of the ODI population.

These findings may be unique to the articles indexed by the PubMed Central database and may not be applicable to research publications in other fields.

Finally, the development of the coding schema and the actual coding of the articles were performed by the same researchers, which could have artificially inflated the agreement rate.

Future Studies

This study only scratches the surface of what we can learn and understand about the Opportunistic Discovery of Information through scientific publications. As mentioned in the limitations, the sampling method used on this study focused on the most recent publications using each term, and therefore, only provides a snapshot of ODI over the last decade or so. Our research team is currently working to characterize ODI within PubMed Central in its entirety, by employing stratified random sampling to more accurately represent the 2.5 million archived articles. We also suspected, based on the secondary analysis performed in this study, that the use of two-word terms could have inadvertently omitted instances of ODI. Although data analysis is ongoing at the time of this submission, the initial search returns based on single terms, indicate a starting population that is over 23 times greater than the one found in this study. While this immense population may not have the relevance frequencies of this initial study, it foreshadows immense possibilities. The random sampling will allow us to analyze changes in ODI reporting over time. We also plan to characterize the incidence of ODI reporting related to the term's location in the article and the number of serendipity-related words used in the article.

Future plans include a comparative analysis of archives of different literature bases and development of a system to automate the analysis process.

Conclusion

In this study, we have proposed and analyzed search terms for identifying evidence of serendipity in full-text searches of the existing biomedical literature. We demonstrated the effectiveness of full-text searches over the use of MeSH terms, and proposed a typology for classifying serendipity-related term use. Our specific conclusions are:

1. Evidence of the Opportunistic Discovery of Information exists within the recent biomedical literature.
2. Full-text searches provide much greater evidence of ODI than those identified through the "serendipity" MeSH term.
3. The use of serendipity related terms within the full-text literature can be classified as one of ten types, four of which are useful for identifying ODI experiences.
4. The single most effective search term for returning reports of ODI is "unexpected finding(s)".
5. The frequency of ODI reports in biomedical literature is not miniscule, and these instances constitute a previously untapped resource for expanding our understanding of the Opportunistic Discovery of Information.

References

Badgaiyan, R. D., & Wack, D. (2011). Evidence of dopaminergic processing of executive inhibition. *PloS One*, *6*(12), e28075. doi:10.1371/journal.pone.0028075

Barnett, A. H. (2011). Diabetes-science, serendipity and common sense. *Diabetic Medicine*, *28*(11), 1289–1299.

Barreiro, C., Martín, J. F., & García-Estrada, C. (2012). Proteomics shows new faces for the old penicillin producer Penicillium chrysogenum. *Journal of Biomedicine & Biotechnology*, *2012*, 105109. doi:10.1155/2012/105109

Buechel, H. M., Popovic, J., Searcy, J. L., Porter, N. M., Thibault, O., & Blalock, E. M. (2011). Deep sleep and parietal cortex gene expression changes are related to cognitive deficits with age. *PloS One*, *6*(4), e18387. doi:10.1371/journal.pone.0018387

Case, D. O. (2012). *Looking for Information : a survey of research on information seeking, needs and behavior*. Bingley, UK: Emerald Group Pub. ;

Erdelez, S., Basic, J., & Levitov, D. D. (2011). Potential for inclusion of information encountering within information literacy models. *Information Research*, *16*(3).

Erdelez, S. (1999). Information Encountering: It's More Than Just Bumping into Information. *Bulletin of the American Society for Information Science and Technology*, *25*(3), 26–29. doi:10.1002/bult.118

Erdelez, S. (2004). Investigation of information encountering in the controlled research environment. *Information Processing & Management*, *40*(6), 1013–1025. doi:10.1016/j.ipm.2004.02.002

Erdelez, S. (2009). Information Encountering. In *Theories of Information Behavior* (pp.

179–184). Medford, N.J.: Information Today Inc.

Erdelez, S., Marinov, M., & Allen, C. (2012). Knowledge Discovery System for Research Hypothesis Generation from Serendipitous Findings: A Feasibility Study. In *Proceedings - AMIA 2012 Annual Symposium*. Chicago: American Medical Informatics Association.

Hargrave-Thomas, E., Yu, B., & Reynisson, J. (2012). Serendipity in anticancer drug discovery. *World journal of clinical oncology*, *3*(1), 1–6. doi:10.5306/wjco.v3.i1.1

Houge, G., & Hennekam, R. C. M. (2009). Reply to Happle. *European Journal of Human Genetics*, *17*(7), 882. doi:10.1038/ejhg.2009.49

Krippendorff, K. (2004). *Content analysis : an introduction to its methodology*. Thousand Oaks, Calif.: Sage.

Lee, Y.-C. (2011). *Serendipity in scientific discoveries: Some examples in glycosciences* (Vol. 705).

Ligon, B. L. (2004). Penicillin: Its Discovery and Early Development. *Seminars in Pediatric Infectious Diseases*, *15*(1), 52–57.

Mayor, S. (2010). Serendipity and cell biology. *Molecular biology of the cell*, *21*(22), 3807–3808.

Meyer-Rochow, V. B. (2009). Food taboos: their origins and purposes. *Journal of Ethnobiology and Ethnomedicine*, *5*, 18. doi:10.1186/1746-4269-5-18

Miwa, M., Egusa, Y., Saito, H., Takaku, M., Terai, H., & Kando, N. (2011). A method to capture information encountering embedded in exploratory web searches. *Information Research*, *16*(3).

Mould, R. F. (1995). Invited review: Rontgen and the discovery of X-rays. *British*

Journal of Radiology, *68*(815), 1145–1176.

Patten, M. L. (2005). *Understanding research methods: an overview of the essentials*. Glendale, Calif.: Pyrczak Pub.

Rubanyi, G. M. (2011). The discovery of endothelin: The power of bioassay and the role of serendipity in the discovery of endothelium-derived vasocative substances. *Pharmacological Research*, *63*(6), 448–454.

Smith, C. E. (2011). Geospatial encountering: Opportunistic information discovery in web-based GIS environments. *Proceedings of the ASIST Annual Meeting*, *48*.

Sundarakumar, D. K. (2011). Non-destructive testing in DNB/MD examination. *The Indian Journal of Radiology & Imaging*, *21*(3), 239–240. doi:10.4103/0971-3026.85379

Tluczek, A., Clark, R., McKechnie, A. C., Orland, K. M., & Brown, R. L. (2010). Task-Oriented and Bottle Feeding Adversely Affect the Quality of Mother-Infant Interactions Following Abnormal Newborn Screens. *Journal of developmental and behavioral pediatrics : JDBP*, *31*(5), 414–426. doi:10.1097/DBP.0b013e3181dd5049

Wang, X., Erdelez, S., Allen, C., Anderson, B., Cao, H., & Shyu, C.-R. (2011). Medical image describing behavior: A comparison between an expert and novice. In *ACM International Conference Proceeding Series* (pp. 792–793). Seattle, WA.

Yadamsuren, B., & Erdelez, S. (2010). Incidental exposure to online news. *Proceedings of the ASIST Annual Meeting*, *47*.

Young, G., Ashdown, D., Arnold, A., & Subramonian, K. (2008). Serendipity in urology. *BJU International*, *101*(4), 415–416.

Chapter 3 Distraction to Illumination - Mining Biomedical Publications for Serendipity in Research

This chapter was originally published as:

Allen, C. M., & Erdelez, S. (2018). Distraction to Illumination: Mining Biomedical Publications for Serendipity in Research. *Proceedings of the American Society for Information Science and Technology*, *55*, 11–19.

Abstract

As our technological capabilities for filling information needs improve, developers seek to more effective ways to support different aspects of the users' experience. One aspect that is gaining attention as an emerging support area is serendipity. However, supporting serendipity within a recommender system is difficult because the experience is unexpected and, therefore unpredictable. While researchers agree that algorithms to support serendipity need to be able to provide a balance of surprise and value to the end user (Niu & Abbas, 2017), an understanding of how to provide that balance has not yet been realized. Information that could be puzzling or distracting to someone as they go about their research activities may provide the trigger someone else needs to make a serendipitous connection in their research. Reports of serendipitous occurrences in research settings have been identified in research commentaries (Campanario, 1996) and within full-text research articles (Allen, Erdelez, & Marinov, 2013). This paper investigates the feasibility of automating the identification of information encounters in full-text research articles. This study contributes to the

development of algorithms for supporting serendipity in information systems. We identified four variables that are useful for predicting information encounters in 25-35% of the instances. While we should continue to search for additional predictive variables, these findings present a novel approach to undertaking the support of serendipity in information systems.

Introduction

In the history of scientific research, 'serendipitous' events, or 'happy accidents', are known to have played significant roles in the advancement of innumerable fields. From the development of Post-it® notes (Schwager, 2000) and Super-glue® (Raja, 2016) to the discovery of x-rays (Shapiro, 1986) and the development of Viagra (Lesko, 2017), society has benefitted from these 'pleasant surprises'. The process of serendipity connects seemingly unrelated concepts in what can appear to be a highly subjective manner, and relies on unique characteristics of the encountering individual to make these fortunate connections (Iaquinta et al., 2008; Maccatrozzo, van Everdingen, Aroyo, & Schreiber, 2017). When viewed from an information behavior perspective, however, it becomes evident that serendipity involves behaviors that are much more purposeful than mere accidents. Makri and Blandford (2012) note that realization of 'serendipity' requires insight and contemplation to make connections between disparate ideas. The unexpected circumstances must occur in the presence of someone who has the intellectual background and capacity to understand the significance of the event. That individual must possess an understanding of the context in which the unexpected event occurs and must also be able to apply that understanding to a different problem. In addition to being unexpected, the user must deem the information interesting (Ge, Delgado-Battenfeld, &

Jannach, 2010). The individual must both perceive and attend to the unanticipated information for it to be useful in making connections between concepts. These characteristics are defining features of information encountering (Erdelez, 1997). Information encountering is the process of noticing and making note of unexpected information that, while not relevant to the most immediate information need, addresses a background need or an unrealized future information need. Information that is encountered rather than sought presents the user with features that are not only surprising but also are salient in a way that allows the user to make loose associations (Kefalidou & Sharples, 2016) between the presented information and distant concepts. The idea of 'loose associations' is particularly relevant to information encountering as additional, purposeful information search is required to build solid connections between the concepts. These associations are the starting point for making the concrete connections necessary for achieving 'serendipitous' outcomes. Serendipity forms the foundation for linking seemingly unrelated ideas and using that connection to improve either or both processes. At the heart of serendipity is a unique type of information behavior – information encountering. By understanding and identifying the ways in which information encountering is described in published research literature, we may be able to develop better methods of supporting serendipity within our information systems and catalyze the positive outcomes of serendipity.

Serendipity and Information Recommender systems

With the ubiquity of electronic resources, information systems have evolved from mere retrieval of information records to recommender systems which seek to supply end users with the most relevant and salient resources from the vast array of indexed

information. Such recommender systems are powered by carefully developed algorithms (Kefalidou & Sharples, 2016; Wu, He, & Yang, 2012). While most algorithms focus on the accuracy of the retrieved results, there has been increased interest in the development of algorithms that will go beyond accuracy to support novelty, diversity and serendipity with the retrieved results. Development of recommender systems that focus on providing 'serendipitous' encounters have been the focus of several information behavior studies (Erdelez, 2004; Makri, Toms, McCay-Peet, & Blandford, 2011; McCay-Peet & Toms, 2011; Wopereis & Braam, 2018) and human computer interaction studies (Dahroug et al., 2017; Fazeli et al., 2017; Maccatrozzo et al., 2017; Niu & Abbas, 2017) in recent years. One challenge to the development of recommender systems that support serendipity is identifying the type of information that end users will find surprising but also useful. Iaquinta et al. (2008) found that the concept of serendipity in recommendations is highly subjective, and that serendipity as a quality of recommended information is difficult to assess. Many developers (Herlocker, Konstan, Terveen, & Riedl, 2004; Maccatrozzo et al., 2017; Niu & Abbas, 2017) have defined serendipity in returned results as those items which hold both surprise and value for the end user. Ge et al (2010) identified key features of serendipitous results as items which have not been previously discovered or expected by the user and are considered to be interesting, relevant and useful to the user. Other researchers have described serendipity in recommender systems as a user experience rather than a feature of the information presented (McNee, Riedl, & Konstan, 2006). While it has been widely accepted that recommender systems need to support serendipity, the development of a mechanism for identifying which items may be deemed both surprising and valuable remains elusive.

Identifying the Information Encounter

In a qualitative study of full-text research journal articles by Allen et al. (2013), it was found that serendipitous information encounters during the research process can be identified by examining the context of synonyms for surprising research findings. This study revealed that synonyms for serendipity were sometimes used to describe the final fortuitous outcome, i.e. the application of an information encounter to a new context in a beneficial way. Other times, synonyms for serendipity were used to describe an information encounter that had not yet been explored or applied to a different situation. These instances described the recognition of surprising or unexpected data during the research process, and were termed 'mentioned findings'. As recommender systems for the support of serendipity requires identifying information that users will find surprising, these 'mentioned findings' articles that contain information that the author has already described as surprising may provide a mechanism for fulfilling that aspect of the serendipitous experience. This finding bears further investigation.

Rationale and Contribution

This paper investigates features of the ways researchers use terms related to serendipity in their research publications as a mechanism for identifying characteristics within indexed information resources that may be useful for supplying both surprising and valuable search results to recommender system end-users. By performing a quantitative analysis on the relationship between reports of actual information encounters and the synonyms for serendipity used by the author and the location of the report within the research article, we will look for predictive variables that may enable the creation of

algorithms to automate the identification of these reports. Using logistic regression, this quantitative project addresses the research question:

What characteristics of author reporting are useful for predicting references to actual serendipitous events?

At this stage in the reserarch, serendipity will generally be defined as a process or experience beginning with an information encounter which is followed up on by an individual with related knowledge, skills and understanding that allows them to realize a valuable outcome.

Background

Serendipity in research and technological development has been analyzed in numerous and disparate fields, including chemistry (Giesy, Newsted, & Oris, 2013), environmental science (Wilkinson & Weitkamp, 2013), and psychiatry (Siris, 2011), and is a widespread topic of interest. Because it crosses disciplinary boundaries and approaches problems in novel ways, serendipity is a phenomenon that has the potential to connect disparate ideas and produce outcomes that improve our quality of life. McCay-Peet and Toms (2018) have advanced a model of serendipity that consists of 3 major stages (See Figure 1). This model was synthesized from five different conceptions (Corneli, Jordanous, Guckelsberger, Pease, & Colton, 2014; Makri & Blandford, 2012; McCay-Peet & Toms, 2015; Rubin, Burkell, & Quan-Haase, 2011; Sun, Sharples, & Makri, 2011) and provides a framework for

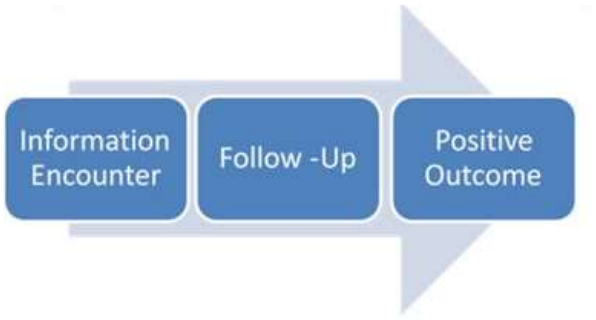

Figure 3.1 Basic Model of Serendipitous Experiences

understanding the experience of serendipity. In this model, a serendipitous experience begins with a triggering event that is unexpected and is coupled with insight by the subject that identifies the encounter as valuable. In order to capitalize on this value, the subject must follow up the finding until positive results are achieved. For this experience to reach the public as serendipitous, the the subject must also go through a period of reflection, where they acknowledge the unexpectedness of the trigger and view the process as fortuitous, describing it to others as such. It is known that disciplinary culture and the expectations of funding bodies may supress reports of serendipity, especially for early-career researchers and those whose observations may not fit with commonly accepted models of understanding. This model of serendipity holds true for those unexpected circumstances that are subsequently explored and their value brought to fruition.

Serendipity can be defined as a fortuitous experience that can occur in any realm of personal or professional occupation. Key aspects of a complete serendipitous experience include information that is acted upon, an element of unexpectedness, and a beneficial outcome. Serendipity is a process that starts with an observation which, when combined with an element of insight, brings about a beneficial outcome (McCay-Peet & Toms, 2015). The unexpectedness may arise at any point in the process. The initial observation may provide unexpected information. An unexpected information encounter may trigger a fortuitous insight. The process of observing and acting upon information may lead to an unexpected outcome. These unexpected aspects of information behavior that occur during the serendipitous experience are information encounters.

Information encountering was first described as a model of information behavior by Sanda Erdelez (1997) and has been analyzed in contexts ranging from news reading (Yadamsuren & Erdelez, 2016) to information literacy (Erdelez, Basic, & Levitov, 2011). Because serendipity is associated with positive outcomes, there is interest in supporting these types of information experiences in multimedia interfaces (McCay-Peet & Toms, 2017), with the expectation that increased information encounters will foster novel and creative leaps of understanding. Some work has been done fostering information encountering in music listening (Wopereis & Braam, 2018) and in multi-media recommender systems (Khalili, van Andel, van den Besselaar, & de Graaf, 2017; Kumpulainen & Kautonen, 2017; Rubin et al., 2011).

Methodology

As a starting point for investigating how biomedical researchers use terms related to serendipity in their research publications, we looked first to published research accounts as indexed in PubMed Central. To identify the representation of serendipity within the language of research articles, we began with a content analysis of full-text articles.

Unit of Analysis

The first step of a content analysis involves the identification of the sampling units, i.e., the synonyms related to the idea of serendipity. Search for these synonyms was undertaken through a process of chaining – identifying the synonyms of serendipity, and then identifying the synonyms of the synonyms until the possibilities were exhausted. Terms identified through this chaining process include: accidental, chance, fortuitous,

happenstance, incidental, serendipitous, serendipity, surprising, unanticipated, unexpected, and unforeseen.

Study Population

The analysis sample was drawn from full text journal articles indexed in PubMed Central. PubMed Central was selected for its extensive collection of over two million full-text, primary research reports. Articles indexed in PubMed Central serve as the population for this study. PubMed Central acts as a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine, currently archiving over 4.3 million articles. In addition to full text searching, PubMed Central offers users the ability to search by Medical Subject Headings, or MeSH terms. MeSH headings are frequently used to narrow literature searches to the user's key ideas. However, the MeSH category designation requires the documents to be classified with these terms by an expert examiner and assigned MeSH terms are limited to 10 to 12 terms per document, so minor ideas contained within the articles are not indexed by the MeSH terms. The current PubMed Central ontology includes the MeSH term "incidental findings," which includes the identified synonyms for serendipity: "incidental finding(s)", "finding(s), incidental", "incidental discovery(ies)", and "discovery(ies), incidental". A search of PubMed Central by usage of this MeSH heading returned only 303 articles. As you can see, while PubMed Central has a MeSH term that is mapped to serendipity, use of the MeSH heading failed to adequately return instances of serendipity, revealing the necessity of full text search for the related terms.

Table 1. Article frequency within full-text articles indexed by PubMed Central by serendipity term use, and associated sample size.

Search Term	Incidence	Sample Size	Search Details
accidental	31669	23	accidental[Body - All Words] AND "research and review articles"[filter]
chance	144141	104	chance[Body - All Words] AND "research and review articles"[filter]
fortuitous	11572	8	fortuitous[Body - All Words] AND "research and review articles"[filter]
happenstance	314	1	happenstance[Body - All Words] AND "research and review articles"[filter]
incidental	20820	15	incidental[Body - All Words] AND "research and review articles"[filter]
serendipit*	3029	2	serendipit*[Body - All Words] AND "research and review articles"[filter]
surprising	205758	149	surprising[Body - All Words] AND "research and review articles"[filter]
unanticipated	8756	6	unanticipated[Body - All Words] AND "research and review articles"[filter]
unexpected	122334	89	unexpected[Body - All Words] AND "research and review articles"[filter]
unforeseen	3895	3	unforeseen[Body - All Words] AND "research and review articles"[filter]
Total	552288	400	

Population Strata

A full text search was performed for each of the identified terms with results filtered to include only research and review articles. The population for this study, consisting of all research or review articles containing a synonym for serendipity in the full-text article, includes 552,288 articles returned with frequencies as indicated in Table 1. Instances of use of the terms serendipitous and serendipity, which were identified as synonyms in the chaining procedure, were combined into a single truncated search of serendipit*.

Random Stratified Sampling

In order to most accurately represent the population of articles, a proportional stratified random sampling method was employed. PASS 11 software was used to estimate the proportion of reported serendipity for the population of 552,288 (total over all strata) with 95% confidence, and a precision (half-width of the interval) of 0.1. A sample of 400 articles is necessary to estimate the true proportions within the population. SAS 9.2 Proc SurveySelect was used to proportionally allocate the sample among the strata for the total sample of 400 articles. Table 1, above, details the proportion of the sample drawn from each strata.

Within each strata, each article was assigned a number from 1 to the total number of articles. Article numbers were assigned based on the default order generated by PubMed Central. SAS was used to randomly identify the designated number of articles for review within each strata according to the assigned article number.

Categorization of Term Use

Term use was coded following the coding scheme described by Allen, et al. (2013), where four categories of term use were of particular interest to our investigation of serendipity. These categories include 'Inspiration' where unexpected information formed the basis for the research design, 'Research Focus' where the entire article is devoted to describing the unexpected information and its evaluation, 'Systematic Reviews' which provide overviews of serendipitous contributions in a field and 'Mentioned Findings' where unexpected or surprising information is presented briefly, but is not the focus of the study. The articles coded into the 'Inspiration' and 'Research Focus' categories typically describe the ways in which the researcher followed up on the

information encounter and applied information gleaned from the encounter to a new context or in a novel way. The 'Systematic Reviews' articles reflect on the impact that serendipitous outcomes have had on a discipline. It is the 'Mentioned Findings' category that is intriguing because it presents surprising information, but investigators have not yet identified salience of the information to a particular problem context. In achieving a serendipitous outcome, the surprising information needs to be connected to researchers with the sagacity necessary to apply that information in a novel way. For example, Roentgen's discovery of x-rays is a commonly noted serendipitous discovery. Most accounts, however, fail to mention that several other scientists experimenting with Crookes Tubes had noted the same glow from phosphoric plates, yet failed to connect the significance of that distant glow to the diagnostic tool that has revolutionized medical diagnostics (Harris, 1995, p. 1). When seeking to populate a recommender system with surprising information, it is the 'Mentioned Findings' category that we feel will prove most useful, as these orphaned findings need to have their salience and application revealed. Accuracy of the coding was evaluated on 10% of the sample, using independent coding by two researchers. Inter-coder reliability was calculated for a subset of 100 articles, returning a KALPHA = .9001, indicating a high level of agreement between the coders. All disagreements were resolved by independently recoding units in disagreement or, failing that, by consensus.

Coding Procedures

PubMed Central's search tool was used to identify all articles with instances serendipity-related term use. Full-text versions of each of those articles were uploaded into NVivo 11 for Windows. Each term was tagged within the articles, then individually

analyzed to determine the meaning conveyed by the term. While the particular synonym for serendipity used to select the article for analysis was noted, all synonyms appearing within the text were coded.

From this examination, 10 categories of term use were identified and 3 of those categories were classified as indicating a serendipitous event, as described by Allen, et al. (2013). The reporting of serendipitous experiences as 'Inspiration' for a research study closely follows current models of the experience of serendipity. The research reported in these instances was inspired by an information encounter and the current study was subsequently undertaken (followed up) resulting in a positive outcome (research findings to report). The 'Research Focus' category reports a traditionally designed study (not arising as follow up to previous observations) where the positive result of the study presents in a way that is completely unexpected. This, likewise, is consistent with models of serendipitous experiences. The third way that terms can be used to indicate a serendipitous experience is as a 'Mentioned Findings'. In this category, the anomalous observations are reported, but they have not yet undergone the follow-up process. It is unknown why follow up is yet to occur. Perhaps the researcher does not have the time or funding to pursue this observation further. Perhaps they do not possess the sagacity to capitalize on the finding. Regardless, this type of reporting has yet to realize the possible positive outcomes. This category, in particular, may present opportunities for information systems to connect such findings with investigators with the education and experiences necessary to bring the process to fruition.

Results

Descriptive statistics

The 400 sampled articles yielded 1,228 unique uses of a synonym for serendipity. Each of those instances was independently coded for term use regardless of the strata for which the article was selected. Following coding of the term use, it was found that 26% f the terms referred to some aspect of serendipity or information encountering and 74% had other semantic meanings (See Figure 2). Of the terms relating serendipitous experiences, 5% indicated using another researcher's unexpected finding as inspiration for their study. Twenty-three percent (23%) focused on describing or further applying serendipitously encountered information. Seventy-one percent (71%) of the terms that were relevant to serendipity fell into the 'Mentioned Findings' category. In these instances, authors reported unexpected or surprising observations made in the course of an investigation, but do not further explore the observation or attempt to apply the observation to another context. None of the articles sampled included terms that fell into the 'Systematic Reviews' category.

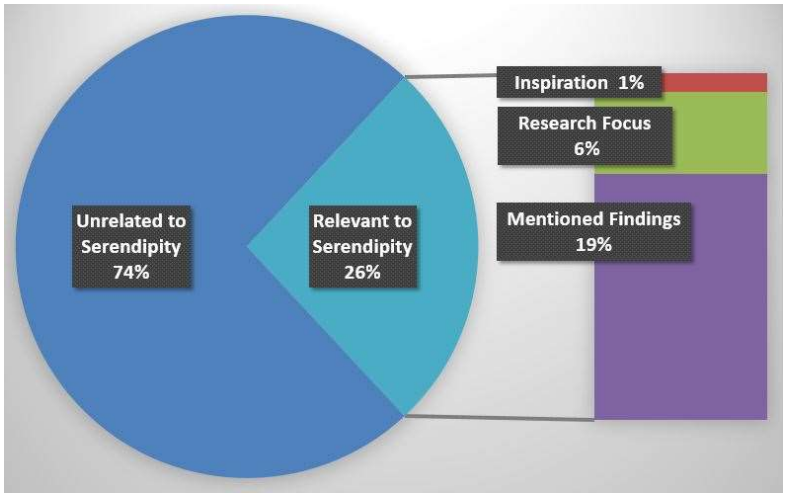

Figure 3.2. Proportion of term use relevant to serendipity or information encountering within the sample.

When extrapolated to the population of 552,288 articles within PubMed containing synonyms for serendipity, over 143,000 articles are estimated to contain

mentions of actual serendipitous events, with nearly 102,000 of those indicating information encounters where the author noted the information as surprising and valuable, but had not yet followed up on the application of that information.

Predictive statistics

Predictive statistics are key to identifying features that will be useful in creating algorithms for a recommender system. Because we have a dichotomous dependent variable (relevant to a serendipitous event or not) and categorical independent variables (serendipity synonym used and location within the article), a binary logistic regression is the preferred predictive analysis. In this analysis, both serendipity synonym used and location of the term within the article were also analyzed for their relationship to actual instances of serendipity. The first step in performing this predictive analysis is to conduct a chi-square analysis to determine if significant relationships exist to support performance of binary logistic regression. In this study the test for overall relationships between search term, location within the article and relationship to actual serendipitous events that was performed was a Fisher's Exact Test. A Fisher's Exact Test is a type of Chi-Square analysis that is performed when one or more cells have a count of 5 or less. The Fisher's Exact Test showed a significant interaction between location and relevance to serendipitous events, with a significance of .000 with a 95% confidence interval. The Fisher's Exact Test on the relationship between synonym used and relevance to serendipitous events also showed a significant interaction, with a significance of .000 with a 95% confidence interval. This test of bivariate relationships was necessary to verify that the data were suitable for performing the binary logistic regression. Because the chi-square test was significant, we know that both variables can be included in the

binary logistic regression. Based on these results from the bivariate tests, we can then choose the variables to include in the final model.

When performing the binary logistic regression, we begin by performing a null analysis, or "null model," where the analysis is performed without including any predictors. The results, when the independent variables are included, are compared to the model when they are not included to see if including the variable significantly impacts the predictive value of the model. For the relationship between synonyms for serendipity and actual serendipitous events, this analysis returned a Wald chi-square statistic of 232.04, which is significant at $p < .05$. This indicates that the coefficients for the variables not in the model are significantly different from zero. This implies that the addition of one or more of these variables to the model will significantly affect its predictive power. The significant search terms in the single model were 'Accidental' and 'Incidental'. Likewise, a null analysis Wald chi-square statistic was performed for the relationship between location within the document and relationship to actual serendipitous events. The residual chi-square statistic for this analysis was 37.804, which is significant at $p < .05$ indicating that the coefficients for the variables not in the model are significantly different from zero. This implies that the addition of one or more of these variables to the model will significantly affect its predictive power. The significant locations in the single model were Abstract and Discussion.

Finally, an overall test of the predicting model was performed. The results of this analysis, which summarizes the relative importance of the explanatory variables individually, can be seen in Table 3. An Enter method was selected and the overall test of the model was statistically significant, indicating that the independent variables in our

model are a significant improvement over those seen in the null model. The independent variables of serendipity synonym and location within the article can accurately predict the whether the author is describing an actual serendipitous event 25-35% of the time. Within that model, the only variables that contribute to this predictive value are the synonyms 'Accidental' and 'Serendipity' and the locations Abstract and Discussion. Location is directly related to the relationship to actual serendipitous events, with terms located in the abstract or discussion sections of the paper being 4 times more likely to refer to actual serendipitous events than terms located in other areas of the article. Interestingly, the relationship between serendipitous events and the synonym used to describe them was actually inverse. The term 'Accidental' was 8.85 times more likely to refer to something other than serendipity; and the term 'Serendipity' was 4.4 times more likely to refer to something else.

Discussion and Conclusions

The descriptive statistics show that there is a sizeable population of articles that contain references to serendipity. The majority of those references are in an early stage, where the researcher noticed unusual or unexpected information in the course of their research and valued that information enough to include it in their research publication. As an information behavior, this triggering episode is termed an information encounter. While information encountering is integral to serendipity, additional factors impact the realization of positive benefits from the information encounter. The large population of 'Mentioned Findings' – type reports of serendipity foreshadow the possibility of using this pool of surprising or unexpected information to support serendipity within

information systems. In order to effectively incorporate these instances into an information system, the identification and cataloging of these reports must be automated.

The results of the binomial logistic regression allowed us to identify four variables that contribute to the effective identification of information encounters. The location of the report within the research article is very useful for predicting whether the author is relating an information encounter. As research articles tend to follow a similar format, automating the identification of these segments of the articles should be fairly straightforward. It was interesting that the terminology used to convey the information encounters was not an effective positive predictor of information encountering reports. Perhaps this is due to the varied scientific fields indexed by PubMed Central. Analysis of term use as it relates to serendipity and information encountering in a narrow disciplinary field is necessary to understanding the linguistic indicators of serendipity.

It was most interesting that 'serendipity' was negatively correlated with references to an active serendipitous experience. Perhaps this is due to the reflective nature of serendipity. It is possible that it is not until after the experience is seasoned and some perspective is gained that the experience is classified as serendipitous. In the midst of experiencing the unexpected, authors experience the confusion and distraction, but have not yet achieved illumination. Overall, this study contributes to the information sciences and human computer interaction literature by identifying a novel way for identifying items that users of information systems may find both surprising and valuable. Furthermore, analysis of ways in which authors represent these information encounters in their research journal articles provides predictive evidence for the development of recommender algorithms.

Challenges and Limitations

While the current study identified two significant variables to predicting reports of serendipity, these variables only account for 25-35% of the variability in the reports of serendipity. Further research is needed to identify additional characteristics that can facilitate the identification of these reports. Initial qualitative analysis hints at the possible existence of domain-specific terminology that could account for greater proportions of the variability in the reports of serendipity. Furthermore, as PubMed Central indexes multiple academic disciplines within the biomedical spectrum, it is possible that the variables identified in this study account for a much greater percentage of the variability in some fields, but are not significant at all in others. In depth qualitative analysis of articles with identified reports of serendipity could prove fruitful.

While this quantitative model holds true for the articles indexed in PubMed Central, it may not be applicable to fields not indexed by PubMed Central. Furthermore, the cross-section of disciplines within PubMed Central may have masked predictive variables by combining competing disciplinary languages. For example, 'accidental' findings that indicate serendipity in a field like chemistry or genetics may be hidden by the use of 'accidental' to discuss traumatic injury in fields like orthopedics, pediatrics, or rehabilitation sciences.

Another challenge to populating information systems with information encounters reported in full-text journal articles is the limited accessibility to the published content. This study was performed using PubMed Central because it provides a large archive of freely accessible full text research articles. Accessing articles indexed through subscription services may prove cost prohibitive and inhibit access to information

encounters in fields that are primarily indexed in subscription-based archives. Furthermore, PubMed Central already incorporates an option for full-text search within their user interface. Identifying the articles for initial screening is much more challenging without the option for full text search. Comparison of article identification through archives that are equipped for full-text search to the secondary article identification process that would be necessary to analyzing the contents of archives without a full-text search feature could be warranted.

Possibilities and Future Directions

In addition to connecting researchers to unexpected information, information systems that support serendipity may be helpful in identifying interdisciplinary collaborators. But research funding mechanisms, and thus the landscape of the research arena, are changing (Aagaard, 2017; Whitley, 2018; Kenney, 2017). Successful research programs are expected to bring in $5-8 for every dollar invested in their infrastructure. Such returns on investment are impossible to achieve by independent researchers bringing in single project grants. In an effort to maintain economic viability, research programs are increasingly moving away from the approach of research as a solitary, isolated pursuit, to an interdisciplinary, center-based model that is underpinned by computational informatics (Goecks, Nekrutenko, & Taylor, 2010; Myers et al., 2004) and leverages academic and industry resources to find solutions for complex, ill-defined problems (Bozeman, 2004; Whitley, 2018). They require a team of researchers pooling their efforts around a collective problem, sharing resources and building synergy, where they accomplish more together than any of them could individually. Building economically viable research teams requires the development of collaborations across

disciplinary, organizational and institutional boundaries. Institutions currently approach the development of these teams rather haphazardly. Holding conferences can provide opportunities to develop these serendipitous connections, but success is far from guaranteed. In fact, many of the most successful research teams note that their partnerships seem to arise serendipitously. Information systems that mine and report information encounters would connect users not only to new, surprising and valuable information, but also to new researchers from diverse fields whose perspectives could enhance and amplify research productivity. By focusing on the information behaviors involved in serendipity, we can capitalize on the observations of others, connecting those observations and illuminating connections for the individuals who are best suited to bring about positive outcomes from them.

References

Allen, C., Erdelez, S., & Marinov, M. (2013). Looking for Opportunistic Discovery of Information in Recent Biomedical Research – A Content Analysis. In *Proceedings of the 76th ASIS&T Annual Meeting* (Vol. 49). Montreal: Richard B. Hill.

Campanario, J. M. (1996). UsingCitation Classics to study the incidence of serendipity in scientific discovery. *Scientometrics*, *37*(1), 3–24. https://doi.org/10.1007/BF02093482

Corneli, J., Jordanous, A., Guckelsberger, C., Pease, A., & Colton, S. (2014). Modelling serendipity in a computational context. *ArXiv:1411.0440 [Cs]*. Retrieved from http://arxiv.org/abs/1411.0440

Dahroug, A., López-Nores, M., Pazos-Arias, J. J., González-Soutelo, S., Reboreda-Morillo, S. M., & Antoniou, A. (2017). Exploiting relevant dates to promote serendipity and situational curiosity in cultural heritage experiences. In *2017 12th*

International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP) (pp. 84–88). https://doi.org/10.1109/SMAP.2017.8022673

Erdelez, S. (1997). Information encountering: a conceptual framework for accidental information discovery (pp. 412–421). Presented at the ISIC '96: Proceedings of an international conference on Information seeking in context, Taylor Graham Publishing. Retrieved from http://portal.acm.org/citation.cfm?id=267217

Erdelez, S. (2004). Investigation of information encountering in the controlled research environment. *Information Processing & Management*, *40*(6), 1013–1025. https://doi.org/10.1016/j.ipm.2004.02.002

Erdelez, S., Basic, J., & Levitov, D. D. (2011). Potential for inclusion of information encountering within information literacy models. *Information Research*, *16*(3).

Fazeli, S., Drachsler, H., Bitter-Rijpkema, M., Brouns, F., Vegt, W. van der, & Sloep, P. B. (2017). User-centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg. *IEEE Transactions on Learning Technologies*, *PP*(99), 1–1. https://doi.org/10.1109/TLT.2017.2732349

Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: evaluating recommender systems by coverage and serendipity (p. 257). ACM Press. https://doi.org/10.1145/1864708.1864761

Giesy, J., Newsted, J., & Oris, J. (2013). Photo-enhanced toxicity: serendipity of a prepared mind and flexible program management. *Environmental Toxicology and Chemistry*, *32*(5), 969–971. https://doi.org/10.1002/etc.2211

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life

sciences. *Genome Biology*, *11*, R86. https://doi.org/10.1186/gb-2010-11-8-r86

Harris, E. L. (1995). *The shadowmakers: A history of radiologic technology* (1st edition). Albuquerque, N.M: American Society of Radiologic Technologists.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. https://doi.org/10.1145/963770.963772

Iaquinta, L., Gemmis, M. d, Lops, P., Semeraro, G., Filannino, M., & Molino, P. (2008). Introducing Serendipity in a Content-Based Recommender System. In *2008 Eighth International Conference on Hybrid Intelligent Systems* (pp. 168–173). https://doi.org/10.1109/HIS.2008.25

Kefalidou, G., & Sharples, S. (2016). Encouraging serendipity in research: Designing technologies to support connection-making. *International Journal of Human-Computer Studies*, *89*, 1–23. https://doi.org/10.1016/j.ijhcs.2016.01.003

Khalili, A., van Andel, P., van den Besselaar, P., & de Graaf, K. A. (2017). Fostering Serendipitous Knowledge Discovery Using an Adaptive Multigraph-based Faceted Browser. In *Proceedings of the Knowledge Capture Conference* (pp. 15:1–15:4). New York, NY, USA: ACM. https://doi.org/10.1145/3148011.3148037

Kumpulainen, S. W., & Kautonen, H. (2017). Accidentally Successful Searching: Users' Perceptions of a Digital Library (pp. 257–260). ACM Press. https://doi.org/10.1145/3020165.3022124

Lesko, L. (2017). Efficacy From Strange Sources. *Clinical Pharmacology & Therapeutics*, *103*(2), 253–261. https://doi.org/10.1002/cpt.916

Maccatrozzo, V., van Everdingen, E., Aroyo, L., & Schreiber, G. (2017). Everybody,

More or Less , likes Serendipity (pp. 29–34). ACM Press. https://doi.org/10.1145/3099023.3099064

Makri, S., & Blandford, A. (2012). Coming across information serendipitously – Part 1: A process model. *Journal of Documentation*, *68*(5), 684–705. https://doi.org/10.1108/00220411211256030

Makri, S., Toms, E. G., McCay-Peet, L., & Blandford, A. (2011). Encouraging Serendipity in Interactive Systems. In *Human-Computer Interaction – INTERACT 2011* (pp. 728–729). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23768-3_136

McCay-Peet, L., & Toms, E. (2011). Measuring the dimensions of serendipity in digital environments. *Usos y Gratificaciones: La Medición de Las Dimensiones de La Serendipia En Entornos Digitales.*, *16*(3), 6–6.

McCay-Peet, L., & Toms, E. (2015). Investigating serendipity: How it unfolds and what may influence it. *Journal of the Association for Information Science and Technology*, *66*(7), 1463–1476. https://doi.org/10.1002/asi.23273

McCay-Peet, L., & Toms, E. (2017). Researching Serendipity in Digital Information Environments. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *9*(6), i–91. https://doi.org/10.2200/S00790ED1V01Y201707ICR059

McCay-Peet, L., & Toms, E. G. (2018). *Researching serendipity in digital information environments*. San Rafael, California: Morgan & Claypool Publishers.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (pp. 1097–1101). New York, NY, USA:

ACM. https://doi.org/10.1145/1125451.1125659

Myers, J. D., Allison, T. C., Bittner, S., Didier, B., Frenklach, M., Green, W. H., … Yang, C. (2004). A collaborative informatics infrastructure for multi-scale science. In *Proceedings of the Second International Workshop on Challenges of Large Applications in Distributed Environments, 2004. CLADE 2004.* (pp. 24–33). https://doi.org/10.1109/CLADE.2004.1309089

Niu, X., & Abbas, F. (2017). A Framework for Computational Serendipity. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 360–363). New York, NY, USA: ACM. https://doi.org/10.1145/3099023.3099097

Raja, P. R. (2016). Cyanoacrylate Adhesives: A Critical Review. *Reviews of Adhesion and Adhesives*, *4*(4), 398–416. https://doi.org/10.7569/RAA.2016.097315

Rubin, V. L., Burkell, J., & Quan-Haase, A. (2011). Facets of serendipity in everyday chance encounters: a grounded theory approach to blog analysis. *Information Research*, *16*(3), 27–27.

Schwager, E. (2000). Little-known facts about little-known people. *Drug News and Perspectives*, *13*(2), 126–128.

Shapiro, G. (1986). *A skeleton in the darkroom: stories of serendipity in science* (1st ed). San Francisco: Harper & Row.

Siris, S. G. (2011). Searching For Serendipity. *The Journal of Clinical Psychiatry*, *72*(8), 1156–1157. https://doi.org/10.4088/JCP.11lr07077

Sun, X., Sharples, S., & Makri, S. (2011). A user-centred mobile diary study approach to understanding serendipity in information research. *INFORMATION RESEARCH-AN*

INTERNATIONAL ELECTRONIC JOURNAL, *16*(3). Retrieved from http://informationr.net/ir/16-3/paper492.html

Wilkinson, C., & Weitkamp, E. (2013). A Case Study in Serendipity: Environmental Researchers Use of Traditional and Social Media for Dissemination. *PLOS ONE*, *8*(12), e84339. https://doi.org/10.1371/journal.pone.0084339

Wopereis, I., & Braam, M. (2018). Seeking Serendipity: The Art of Finding the Unsought in Professional Music. In S. Kurbanoğlu, J. Boustany, S. Špiranec, E. Grassian, D. Mizrachi, & L. Roy (Eds.), *Information Literacy in the Workplace* (pp. 503–512). Springer International Publishing.

Wu, W., He, L., & Yang, J. (2012). Evaluating recommender systems. In *Seventh International Conference on Digital Information Management (ICDIM 2012)* (pp. 56–61). https://doi.org/10.1109/ICDIM.2012.6360092

Yadamsuren, B., & Erdelez, S. (2016). Incidental Exposure to Online News. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *8*(5), i–73. https://doi.org/10.2200/S00744ED1V01Y201611ICR054

Chapter 4 Characterizing Serendipity in Biomedical Research Literature - Identification of Syntactic and Linguistic Features

This chapter was prepared for submission to the Journal of the Association of Information Science and Technology.

Abstract

Electronic information systems play a key role in supplying researchers with critical information to support their scientific endeavors. Developers of these systems are continually looking for ways to improve users' information experiences. One element of user support garnering recent attention is the support of serendipity, but, methods for effectively providing that support have not been fully investigated. In this mixed method study, we apply qualitative content analysis, supported by natural language processing, concurrently with statistical analysis to evaluate the predictive value of features of serendipitous findings reported in biomedical research articles. This study investigates the feasibility of automating the identification of these information encounters in full-text research articles, and contributes to the development of algorithms for supporting serendipity in information systems. Three syntactic features of serendipity reports that demonstrate statistical reliability in predicting an author's reference to serendipity are identified and hold promise for the development of predictive algorithms to support serendipity in information recommender systems. While we should continue to search for additional predictive variables, these findings present a novel approach to undertaking the support of serendipity in information systems.

Introduction

Serendipity is a phenomenon that has the potential to connect disparate ideas and produce outcomes that improve our quality of life. If we can develop information systems to effectively support the factors that promote serendipity, we could increase the realization of these improvements. Furthermore, it has been noted that recommender systems, in the support of serendipity, need to supply the end-user with information that is surprising (Ge, Delgado-Battenfeld, & Jannach, 2010). In previous works (Allen & Erdelez, 2018; Allen, Erdelez, & Marinov, 2013), we have seen that instances of serendipity are alluded to in biomedical research publications, and that these reports occur both immediately following the information encounter that can potentially trigger a serendipitous outcome and after the outcome has been realized. These reports were also shown to exist in sufficient numbers to provide the pool of surprising information needed to populate such a recommender system. Yet, there is a need for close examination of the ways in which researchers report their serendipitous experiences. By identifying specific syntactic and semantic features employed by the authors in reporting serendipity, we can better informing the development of algorithms for the retrieval and recommendation of these concepts.

Rationale and Contribution

The purpose of this mixed methods study is to describe the phenomenon of serendipity as it is recorded in biomedical research articles indexed in the PubMed Central database in a way that might inform the development of machine compilation systems for the support of serendipity. Within this study, serendipity is generally defined as a process or experience beginning with encountered information that is analyzed and

further pursued by an individual with related knowledge, skills and understanding that allows them to realize a valuable outcome. The information encounter that initiates serendipitous experience exhibits qualities of unexpectedness as well as value for the user.

Literature Review

The Merriam-Webster dictionary defines serendipity as "the faculty or phenomenon of finding valuable or agreeable things not sought for." Vocabulary.com goes on to describe serendipity as "good luck in making unexpected and fortunate discoveries." Merton and Barber (2004) describe a typical pattern of serendipitous experience as one in which the investigator observes an "unanticipated, anomalous and strategic datum which becomes the occasion for developing a new theory or extending an existing theory." The inconsistency of the observation provokes curiosity. The unexpected observation "bears upon [pertains to; concerns] theories not in question when the research was begun." The inconsistency compels the investigator to make sense of the data, and, further, to relate the data within the broader frame of reference. The implications of the observation become applicable in a generalizable way. Finally, the observer brings something of themselves to the data. The observer must be theoretically sensitized in order to "detect the universal in the particular."

These definitions emphasize the information behavior involved in the experience of serendipity. Within the information sciences, McCay-Peet and Toms (2018) have been instrumental in synthesizing a model of serendipity that unifies the work of many researchers including Corneli (2014), Makri (2012), Rubin (2011), and Sun (2011). This model, much like the observations of Merton and Barber, frames serendipity as a process

that begins with an information encounter. The surprising information is then explored, and used to connect disparate ideas, ultimately resulting in a valuable outcome.

Researching Serendipity

Research on serendipity as an information behavior has increased dramatically over the past 25 years (Erdelez et al., 2016). While some researchers have focused on the features of serendipitous events, information, or individuals (McCay-Peet & Toms, 2018), others have focused on the entire process or experience of serendipity. For example, Erdelez's theory of information encountering effectively describes the event of a serendipitous discovery occurring in an information acquisition context (Erdelez, 2004), where users bump into and attend to information that is unrelated to their current information search, but instead related to a background problem that the seeker has not yet resolved. After noting the usefulness of the encountered material, the seeker stores the information for later use and returns to the original search. Erdelez (1997) has also investigated defining features of individuals related to the frequency of their serendipitous encounters. In this study, it was found that certain personality traits, such as curiosity or varied interests are more highly reported by "super-encounterers" who were serendipity-prone. Across the literature, some common features of the serendipity experience have been identified. The first feature is the existence an information encounter. This information, as noted by Rioux (2005, p. 171), is sufficiently unique, useful or otherwise meaningful to prompt the individual to examine the information more closely. Thus, the information encounter serves as a cognitive trigger to the individual and prompts them to actively attend to the information presented. But the serendipitous outcome is not realized just because an individual encounters anomalous information;

that individual must also possess the "sagacity" to apply that information in a way that enables them to recognize the value of the information. "Sagacity" is a term associated with serendipity from its inception. Meaning "foresight, discernment, or keen perception" ("sagacity," n.d.), Horace Walpole used the idea of "accidental sagacity" in the letter that coined the term "serendipity" (Foster & Ford, 2003). Regardless of the descriptor employed, it has come to be recognized that serendipity cannot exist apart from a prepared mind. That preparation, along with perseverance and an incubation period, are necessary to achieving a valuable outcome, which may be global or individual in scope. In addition to the common features mentioned above, three of the serendipity models (Makri & Blandford, 2012; McCay-Peet & Toms, 2015; Rubin et al., 2011) include a reflection component following the accomplishment of the outcome in which the individual acknowledges the serendipity of the experience and decides to frame it as such. As technological support plays an ever-increasing role in research productivity, the development of systems to support serendipity becomes increasingly important.

Information Recommender Systems and Serendipity

As the proliferation of indexed information continues to grow, so to have the demands on electronic information systems. Information systems have historically focused on supplying end users with improving the accuracy and saliency of the returned information. However, as we have grown more dependent upon electronic retrieval systems, the demand for systems that support additional types of information behaviors has increased (Ge et al., 2010). Users now expect information systems to go beyond mere retrieval of information records to offer recommendation and prioritization of resources as they relate to the user's query. Development of these recommender systems utilizes

carefully developed algorithms (Kefalidou & Sharples, 2016; Wu, He, & Yang, 2012). In recent years, there has been increased interest in algorithms that support novelty, diversity and serendipity in addition to assuring the accuracy of the retrieved results. Investigators in the fields of human computer interaction (Dahroug et al., 2017; Fazeli et al., 2017; Maccatrozzo, van Everdingen, Aroyo, & Schreiber, 2017; Niu & Abbas, 2017) and information Behavior Studies (Erdelez, 2004; Makri, Toms, McCay-Peet, & Blandford, 2011; McCay-Peet & Toms, 2011; Wopereis & Braam, 2018) have both focused on the development of recommender systems that support serendipitous encounters.

The work toward developing support for serendipity through information recommender systems involves the capture of semantic meaning. Developers are faced with the challenge of defining the key features and attributes that characterize serendipity. The key constructs currently recognized in the identification of serendipity include the acquisition of valuable information and the presence of an element of surprise (Herlocker, Konstan, Terveen, & Riedl, 2004; Maccatrozzo et al., 2017; Niu & Abbas, 2017).

Identifying Instances of Serendipity

In a recent study, Allen et al. (2013) found that evidence of serendipity could be identified within scientific research literature by searching for and analyzing synonyms for serendipity. This process holds the potential for conversion to automated text-mining, if features can be identified to distinguish true reports of serendipity from other uses of those words. In that study, 14 sets of two-word terms and their plural variants were used in full-text searches of 436 articles drawn from PubMed Central. Stratified sampling was

used but the articles were drawn from most recent publications rather than being randomly selected. The study followed the two-word text searches with qualitative analysis of each instance of the terms. This analysis revealed six semantic meanings that were irrelevant to the concept of serendipity. Meanings that were deemed irrelevant to the concept of serendipity include statistical references, where the author describes a result as a 'chance finding', or historical references, where the author notes a 'fortuitous discovery' in their field when providing context for the current study. Allen et al. also found four uses that were significant to instances of serendipity. These included 'inspiration', 'research focus', 'mentioned findings', and 'systematic reviews'. In that sample 58% of the examined linguistic cues for serendipity were actually related to reports of serendipitous experience. This study builds on techniques developed in the Allen study by identifying textual features that may be useful in compiling reports of serendipity and evaluating the predictive value of those features.

Methods

This mixed-methods study combines qualitative content analysis with mathematical frequencies that support statistical analysis. In a content analysis a body of documents is examined to reveal both manifested and latent content communicated in the material (Krippendorff, 2004). Details of the design are described below.

Study Population and Corpus Compilation

The first step in undertaking the content analysis is compiling the corpus of documents to be analyzed. According to Gries (2009, p. 7), " 'corpus' refers to a machine-readable collection of spoken or written texts that were produced in a natural communicative setting and compiled with the intention to be 1) representative and

balanced with respect to a particular linguistic variety … and 2) to be analyzed linguistically." In compiling the corpus, we begin by locating a source of machine-readable documents that represent a pool of "natural" communications (as opposed to texts created specifically for computer processing). As our research question focuses on the experience of serendipity within the research context, a population of research articles was needed to address this question. While many disciplines exist within the realm of scientific research, we have chosen to focus this study on biomedical sciences. This focus was chosen for two reasons. First, a large, open-source database indexing biomedical research articles in full text format is available through PubMed Central. Open access to the full text journal articles and the ability to automatically search for specific terms within the entire full text of an article without having to import the article into a natural language processing program expedited completion of the study. Additionally, biomedical sciences include translational and applied sciences that look to improve the health and well-being of society through the application of natural sciences to health and social care. By facilitating serendipitous encounters in the biomedical sciences, we could facilitate the development of new and improved ways of treating and caring for patients. The population for this study consists of all articles that contain words that may signal a report of serendipity and are indexed within the PubMed Central database.

The second step in compiling the corpus is identifying the pool of machine-readable documents that contains sufficient linguistic cues, hereafter termed 'surprise descriptors,' to identify the concept we are interested in studying. In content analysis terminology, the words used as linguistic indicators of a concept are termed 'sampling units.' For this study, the sampling units chosen are single-words that express the surprise

component of a serendipitous experience. These 'surprise descriptors' provide linguistic cues that may indicate presence of a report of serendipity. This set of terms was derived through a method of chaining, where each of the synonyms for serendipity were listed and each of the synonyms for those synonyms were looked up and added to the list until the pool of terms was exhausted. This chaining process resulted in identification of the following surprise descriptor terms: 'accidental', 'chance', 'fortuitous', 'happenstantial', 'incidental', 'serendipitous', 'serendipity', 'surprising', 'unanticipated', 'unexpected', and 'unforeseen'. Instances of use of the terms serendipitous and serendipity, which were identified as synonyms in the chaining procedure, were combined into a single truncated search of serendipit*.

To move from the set of surprise descriptors to the identification of the articles comprising the study population, a full-text search was performed for each of the identified terms using the PubMed Central advanced search builder. The advanced search builder allowed us to search for individual terms within all words in the body of the text, as well as limiting the results to research and review articles. Once the results were filtered to remove any duplicate articles, this full text search returned over 550,000 articles. The frequencies of term use ranged from 314 articles using the word happenstance to 205,000 articles using the term surprising. Details of the frequencies of each sampling unit, as well as the search statements used, can be found in Table 1.

Table 1. Frequencies of Surprise Descriptors

Surprise Descriptor	# of Articles	Percentage	Sample Size	Search Details
accidental	31669	5.73%	24	Accidental[Body – All Words] AND "research and review articles" [filter]
chance	144141	26.10%	104	chance*[Body – All Words] AND "research and review articles" [filter]*
fortuitous	11572	2.10%	8	fortuitous*[Body – All Words] AND "research and review articles" [filter]*
happenstantial	314	0.06%	0	happenstantial*[Body – All Words] AND "research and review articles" [filter]*
incidental	20820	3.77%	15	incidental*[Body – All Words] AND "research and review articles" [filter]*
serendipit*	3029	0.55%	2	serendipit**[Body – All Words] AND "research and review articles" [filter]*
surprising	205758	37.26%	149	surprising*[Body – All Words] AND "research and review articles" [filter]*
unanticipated	8756	1.59%	6	unanticipated *[Body – All Words] AND "research and review articles" [filter]*
unexpected	122334	22.15%	89	unexpected*[Body – All Words] AND "research and review articles" [filter]*
unforeseen	3895	0.71%	3	unforeseen*[Body – All Words] AND "research and review articles" [filter]*
Total	**552,288**	**100%**	**400**	

Now that the group of machine-readable documents containing cues indicative of serendipity has been identified, it is important to select from that population a corpus that is representative of the population as a whole. In order to achieve the most accurate representation of the population of articles, a proportional stratified random sampling method was employed. PASS 11 software was used to calculate the sample size needed to provide sufficient statistical power to make inferences regarding the population. For the population of 552,288 (total over all strata) with 95% confidence, and a precision (half-width of the interval) of 0.1, a sample of 400 articles is necessary to estimate the true proportions within the population. SAS 9.2 was used to proportionally allocate the sample among the strata for the total sample of 400 articles. Table 1 also details the proportion of the sample drawn from each strata. In order to obtain the random sample, a

list of articles was compiled for each context strata. Each article was assigned a number based on the default order in which it was presented by the PubMed Central database. SAS 9.2 was used to identify, by article number, which articles were included in the sample. Once the sample articles were determined, PDFs of each article were obtained and uploaded into NVivo 11 for Windows to facilitate the qualitative analysis.

Coding Procedures

Once the PDFs were uploaded into NVivo 11, a text search query was performed to locate each of the linguistic indicators. All terms used in the initial retrieval of the articles were identified and marked for coding regardless of the term for which the article was included in the study. This resulted in 1162 unique passages being marked for analysis. The text query for each linguistic indicator was saved as a separate node within the NVivo project. At this point the project files were made available to the coders. To demonstrate the clarity of the coding scheme, coding was performed individually by two trained coders who were not involved in the creation of the initial coding scheme. Each coder analyzed the paragraph containing the linguistic indicator and coded the way in which the term was used as well as its location within the article.

In this study, we began by coding the meaning imbued by each surprise descriptor using the coding scheme published by Allen, at al (2013). Additional categories of term use were added as necessary when instances did not fit with the published codes. In addition to coding the semantic meaning of each term, coders were asked to code the location of the term within the research document. The location of the term provides information about the context in which the serendipitous experience occurred and possibly informs the automated compiling of serendipitous experiences. According to

Kapetanios (2013, p. 145), the location can be classified as a property in the development of query algorithms. These properties can be used in the development of models that represent the meaning of terms. Since the corpus of documents did not include location information for the terms in the documents, these needed to be manually coded by our researchers. The locations used in our coding include 'title', 'abstract', 'introduction', 'literature review / background', 'methods', 'results', 'discussion', and 'conclusion'. While some articles did not explicitly use the headings listed above, the coders analyzed both the location and the context of the mention to code it into the traditionally-labeled areas. For example, if the surprise descriptor occurred within the first 3-4 paragraphs of the article, that was considered to be congruous to the introduction. Surprise descriptors occurring after the first heading but prior to descriptions of methods were classified into the literature review category.

Accuracy of the coding was evaluated on 10% of the sample, using independent coding by two researchers. Inter-coder reliability was calculated for a subset of 100 articles, returning a KALPHA = .9001, indicating a high level of agreement between the coders. All disagreements were resolved by independently recoding units in disagreement or, failing that, by consensus.

Results and Discussion

Incidence of Serendipity within the Biomedical Population

In exploring the semantic meaning communicated with terms related to the idea of surprise, this study tested the null hypothesis that the use of terms related to surprise are not indicative of reports of serendipity. Our results refute this hypothesis, and reveal that serendipity is a fairly common information experience for biomedical researchers.

Surprise descriptor terms were identified in the full-text of 552,288 of the 4.3 million biomedical research articles indexed by PubMed Central. After analyzing the semantic meaning conveyed by the surprise descriptor terms in our sample, we discovered that 41% of articles containing a surprise descriptor term communicate reports of serendipity. When these findings are extrapolated to the population of 552,288 biomedical research articles indexed by PubMed Central containing the surprise descriptor terms, we estimate nearly 225,000 articles, or approximately 7.5% of all articles indexed by PubMed Central, contain a reference to serendipity. This sizable population forms a suitable pool for automatic compilation.

Syntactic Features of Serendipity

In order to begin development of algorithms for identifying these instances of serendipity, we need to identify the syntactic features of the authors' descriptions that could be identified by computing devices. We begin this analysis with the surprise descriptors that we identified as representing the idea of serendipity. By using single terms instead of two terms connected as a single phrase as demonstrated in the Allen study (2013), the search is able to capture reports of serendipity where the surprise descriptor is not immediately connected to the type of information encountered. It also allow for the capture of a wider range of information descriptors than were accounted for in the Allen (2013) study. The effectiveness of this strategy is seen in the dramatic increase in the number of articles retrieved in the search. 552,288 articles were retrieved with the single term search in contrast to the 23,571 articles retrieved in the Allen (2013) study.

As individual words, these surprise descriptors are strings of symbols that can be identified by a computer. Existing algorithms would also allow us to use wildcards to search for other stemmed uses of the terms, i.e. adverbial forms of the same descriptor, such as 'accidentally' or 'surprisingly'. The number of surprise descriptor terms analyzed, as well as the number of instances of those terms that were identified as conveying semantic meaning related to serendipity, are presented in Table 2. The far right column identifies the proportion of instances that were related to serendipitous experiences. We found that the highest proportion of terms related to serendipity were returned with the terms unanticipated and unexpected. This mirrors the finding of the Allen (2013) study.

Table 2. Surprise Descriptor vs. Relevance to Serendipity Cross Tabulation

Surprise Descriptor	Terms Analyzed	# Relevant	% Related to Serendipitous Experiences
Accidental	89	11	12%
Chance	268	0	0%
Fortuitous	24	9	38%
Happenstance	2	1	50%
Incidental	94	21	22%
Serendip*	10	4	40%
Surprising	385	179	46%
Unanticipated	11	6	55%
Unexpected	277	139	50%
Unforseen	2	0	0%
Totals	**1162**	**370**	**32%**

Another syntactic feature that might be helpful in understanding the experience of serendipity is the location within the research article where the description occurs. Analysis of the location within the article allows us to gain insight into the phases of the research process that provide an environment for serendipitous experiences. In this study, over half (53.4%) of the terms coded were located in the discussion section of the

articles. This was followed by the results section (20%) and introduction (2.4%). When location was considered in relation to the relevance of the surprise descriptor, the abstract and discussion sections contained the highest proportion of serendipity relevant cues.

Computing systems are also capable of identifying grammatical structures. As we have noted in the literature review, serendipitous experiences have two defining features. The first is an element of unexpectedness or surprise. The second defining feature of serendipity is the provision of valuable information. In analyzing the syntactic structure of references to serendipity, we found that references to serendipity typically consist of two word phrases that typically act as the object of the sentence, and take the form of an adjective followed by a noun. The adjective typically describes the unexpectedness or surprise element of the serendipity experience. The noun expresses the type of information encountered. This structure is alluded to in the two-word search terms employed in the Allen (2013) study. In the current study, this type of noun phrase was found in 54% of the surprise descriptors analyzed and in 63% of the cues that were found to be related to an actual serendipitous experience.

Predicting Serendipity

In order to create algorithms capable of identifying references to serendipity, we must first determine which features are most predictive of references to serendipity. In this study, we performed a factorial logistic regression to develop a predictive model for identifying instances of serendipity. Factorial logistic regression was selected because of the categorical nature of both the independent variables and the dependent variable. The logistic regression determines the probability of a given surprise descriptor being related to actual serendipitous experiences. The independent variables considered are the chosen

surprise descriptor, the location of the cue within the article, and the presence of a selected information term. The dependent variable is relevance to serendipity. The surprise descriptor variable contains ten (10) different categories, and the location variable consists of 8 different categories. Details regarding these categories are described in the methods section above. The information term variable contains 10 categories. To derive these information terms, a word frequency query was performed in NVivo 12. This query returned the 1000 most frequently used words within the sample articles. Natural language processing was used to identify the terms and synonyms for those terms to arrive at the frequencies. The results of the word frequency were reviewed to identify terms that were related to the idea of information obtained during the performance of a research endeavor. This query produced the relevant terms '*finding*', '*discovery*', '*development*', '*effect*', '*interaction*', '*result*', '*mechanism*', '*information*', '*variation*', and '*evidence*'. A compound query was then performed to locate all instances of these information terms that occurred within 10 words of the previously coded surprise descriptor terms. As NVivo is designed as a qualitative analysis tool, the results of this query could not be exported in a way that documented which surprise terms (and their associated coding regarding relevance to serendipity) were collocated with the identified information terms. Consequently, results from the query had to be compiled by hand in preparation for statistical analysis.

Effect of Surprise Descriptor on Relevance to Serendipity

The data was uploaded into SPSS Statistics 25 for analysis. A total of 1,162 surprise descriptors were analyzed for relevance to serendipity. Thirty- two percent (32%) of the surprise descriptors were coded as relevant to a serendipitous experience.

Within the relevant terms, the surprise descriptor terms ‘surprising’ (48.0%) and ‘unexpected’ (35.7%) occurred most frequently. The Fisher’s Exact Test was used to screen the data for bivariate relationships instead of Pearson Chi-Square because there were cells with less than 5 counts. The Fisher’s Exact Test showed a significant relationship between surprise descriptor terms and relevance to serendipity, p =.003. Regression results for variables not in the model show the residual chi-square statistic to be 232.04, which is significant at $p < .05$. The coefficients for the variables not in the model are significantly different from zero, implying that the addition of one or more of these variables to the model will significantly affect its predictive power. Regression results indicated that that the overall model fit of surprise descriptors as predictors was questionable, (-2 Log Likelihood = 1143.411) but was statistically reliable, in distinguishing relevance to serendipity; X2 (1) = 310.647, p <.000. The model correctly classified 68.4% of the cases. Regression coefficients are presented in Table 3. Wald statistics indicated that, for surprise descriptors, only the terms ‘accidental’ and ‘serendipity’ were significant in predicting a statement’s relevance to serendipity.

Table 2: Regression Coefficients for Surprise Descriptor

Variables in the Equation	B	S.E.	Wald	df	Sig.	95% C.I. for EXP(B)	
						Lower	Upper
Accidental			51.226	10	.000		
Unexpected	2.141	.686	9.746	1	.002	8.509	2.219
Unforseen	.175	.617	.080	1	.777	1.191	.355
Chance	21.385	28420.721	.000	1	.999	1938569811.422	.000
Fortuitous	21.385	2455.176	.000	1	.993	1938569811.422	.000
Happenstance	.693	.738	.882	1	.348	2.000	.471
Incidental	.182	1.538	.014	1	.906	1.200	.059
Serentipitous	1.428	.654	4.766	1	.029	4.171	1.157
Serendipity	.875	.931	.884	1	.347	2.400	.387
Surprising	-21.021	40192.970	.000	1	1.000	.000	.000
Unanticipated	.323	.614	.276	1	.599	1.381	.414
Constant	-1.253	.802	2.441	1	.118	.286	

Interestingly, the relationship between serendipitous events and the synonym used to describe them was actually inverse. The term 'Accidental' was 8.85 times more likely to refer to something other than serendipity. Within the biomedical literature, the term 'accidental' is frequently used to describe trauma situations or other sorts of injuries. It is possible that the statistics on the term 'accidental' in this study were skewed by a few articles within the sample that had a large number of irrelevant terms. For example, one article on toxic exposures (Ponampalam et al., 2009) contained 39 instances of the term accidental; none of these were relevant to a serendipitous experience. Two other articles, one on accidental injury (Mayes & Macleod, 1999) and one on child deaths from injury (Edwards, Roberts, Green, & Lutchmun, 2006) contained 8 and 9 irrelevant instances of the term 'accidental', respectively. The term 'Serendipity' was 4.4 times more likely to refer to something else. In the 10 instances of the term 'serendipity', most of the authors are describing serendipity as a means of classifying something. For example, Rang (1972) states, "The inverted *serendipity* variant. *Serendipity* is making a discovery by chance." Overall, while the surprise descriptors are statistically reliable for identifying cases of serendipity, as a solitary indicator they don't provide a very effective means of automatically mining for the concept. Additional linguistic indicators are needed in order to build an effective mining algorithm. Next, we considered the predictive value of term location on relevance to serendipity.

Effect of Term Location on Relevance to Serendipity

The relationship between the location in the article where the surprise descriptor was identified and relevance to serendipity was also examined. A total of 1100 of the originally identified surprise descriptors existed within the identified article locations.

Within the relevant instances, the locations 'results' (21.7%) and 'discussion' (54.4%) occurred most frequently. The actual frequency counts as well as the percentage of relevant and irrelevant instances within that location in the article are displayed in Table 4.

Table 3. Location of Surprise Descriptor vs. Relevance to Serendipity

	Term relevance				
	Relevant		Irrelevant		
Location in Article	Count	% within Location in Article	Count	% within Location in Article	Total
Title	7	58.3%	5	41.7%	12
Abstract	19	48.7%	20	51.3%	39
Introduction	39	27.1%	105	72.9%	144
Methods	5	7.9%	58	92.1%	63
Literature Review/ Background	5	19.2%	21	80.8%	26
Results	94	40.5%	138	59.5%	232
Discussion	195	31.4%	426	68.6%	621
Conclusion	6	24.0%	19	76.0%	25
Total	**370**	**31.8%**	**792**	**68.2%**	**1162**

The Fisher's Exact Test was used to screen the data for bivariate relationships instead of Pearson Chi-Square because there were cells with less than 5 counts. The Fisher's Exact Test showed a significant relationship between surprise descriptor terms and relevance to serendipity, p =.001. Regression results indicated that the overall model fit of information terms as predictors was questionable (-2 Log Likelihood = 953.287) but was statistically reliable in distinguishing relevance to serendipity; X2 (1) = 27.145, p >.000. The model correctly classified 83.6% of the cases. Regression coefficients are presented in Table 5. Wald statistics indicated that the locations 'abstract' and 'results' significantly predict relevance to serendipity.

Table 4. Regression Coefficients for Location

Variables in the Equation	B	S.E.	Wald	df	Sig.	95% C.I. for EXP(B)	
						Lower	Upper
Abstract			17.289	7	.016		
Conclusion	.519	.875	.351	1	.553	1.680	.302
Discussion	-.999	1.093	.834	1	.361	.368	.043
Introduction	-.381	.809	.221	1	.638	.683	.140
Lit. Review Background	-.429	.832	.266	1	.606	.651	.127
Methods	-1.350	1.086	1.544	1	.214	.259	.031
Results	-2.874	1.288	4.980	1	.026	.056	.005
Title	-.088	.822	.011	1	.915	.916	.183
Constant	-1.253	.802	2.441	1	.118	.286	

Overall, term location was found to be a statistically reliable indicator of serendipity. Although the machine classification of the concept was quite accurate, many instances of serendipity were left unidentified. Other linguistic indicators are still needed to improve the effectiveness of machine compiling. Next, we examined the information term that the surprise descriptor modifies.

Effect of Information Term on Relevance to Serendipity

Because the idea of serendipity is frequently expressed as a noun phrase with an object that represents the type of information encountered in addition to the adjective that describes the surprising element of the discovery, we analyzed the documents for the presence or absence of an information term collocated within 10 words of the surprise descriptor terms that were coded originally. A total of 612 of the originally identified surprise descriptors were found to be collocated with an information term. Thirty-three percent of the information terms were linked to surprise descriptor s that were coded as relevant to a serendipitous experience. The information terms ‘finding’ (25.9%) and ‘result’ (33.2%) occurred most frequently. The Pearson Chi-Square test showed a significant relationship between presence of an information term near a surprise

descriptor and relevance to serendipity, X^2 (9, N = 612) = 24.210, p =.004. Regression results indicated that the overall model fit of information terms as predictors was questionable (-2 Log Likelihood = 753.918) but was statistically reliable in distinguishing relevance to serendipity; X^2 (1) = 26.554, p =.002. The model correctly classified only 66.5% of the cases. Regression coefficients are presented in Table 6. Wald statistics indicated that the information terms 'development', 'variation', 'discovery', 'evidence', 'finding', and 'interaction' significantly predict relevance to serendipity. However, odds ratios for these variables indicated little change in the likelihood of a surprise descriptor indicating a serendipitous experience.

Table 5. Regression Coefficients for Information Terms

Variables in the Equation	B	S.E.	Wald	df	Sig.	95% C.I. for EXP(B)	
						Lower	Upper
Development			21.523	9	.011		
Variation	-.777	.334	5.408	1	.020	.239	.885
Discovery	-1.905	.765	6.203	1	.013	.033	.666
Effect	-.458	.526	.760	1	.383	.226	1.772
Evidence	-.623	.278	5.021	1	.025	.311	.925
Finding	-.826	.419	3.879	1	.049	.192	.996
Information	-.420	.235	3.205	1	.073	.415	1.041
Interaction	-2.163	.756	8.183	1	.004	.026	.506
Mechanism	-.416	.391	1.129	1	.288	.306	1.421
Result	.235	.499	.222	1	.638	.476	3.360
Constant	-.235	.162	2.094	1	.148		

Limitations and Future Research

Overall, significant predictive value was found in all of the independent variables analyzed. However, the overall model fit for all factors was questionable and returned only a small proportion of the identified instances of serendipity. Further research is needed to identify additional features that may be useful in compiling these reports. Additionally, it is possible that the variables identified in this study are not representative of all domains indexed by PubMed Central. This could result in the predictive value of some variables being

inadvertently masked, as relevant term use may exist in greater proportions in one discipline while the same term does not return references to serendipity in another discipline at all. For example, 'accidental' findings may accurately indicate serendipity in fields like genetics or pharmacology, but that effect may be hidden by irrelevant uses of the term 'accidental" in fields that discuss traumatic injury. Furthermore, this qualitative model is applicable to PubMed Central but may not reflect the disciplines indexed by other literature databases. While the statistical model may not capture a large percentage of the existing references to serendipity, instances that are identified are classified fairly accurately. Directions for future study include analyzing the instances for the existence of domain-specific terminology, as well as investigating ways to filter the pool of articles containing reports of serendipity to target the end-users' areas of expertise.

Conclusion

Overall, this study contributes to the information sciences and human computer interaction literature by identifying a novel way for identifying novel observations that users of information systems may find both surprising and valuable. Furthermore, analysis of ways in which authors represent these information encounters in their research journal articles provides predictive evidence for the development of recommender algorithms. This study analyzed the value of three syntactic features of author's descriptions as predictors of serendipitous experiences. All of these features demonstrated statistical reliability for predicting the presence of a description of serendipity. While the fit of the model was less than ideal, it provides a basis for the development of predictive algorithms to underpin information recommendation systems.

References

Allen, C. M., & Erdelez, S. (2018). Distraction to Illumination: Mining Biomedical Publications for Serendipity in Research. *Proceedings of the American Society for Information Science and Technology*, *55*, 11–19.

Allen, C. M., Erdelez, S., & Marinov, M. (2013). Looking for opportunistic discovery of information in recent biomedical research – a content analysis. *Proceedings of the American Society for Information Science and Technology*, *50*, 1–11.

Corneli, J., Jordanous, A., Guckelsberger, C., Pease, A., & Colton, S. (2014). Modelling serendipity in a computational context. *ArXiv:1411.0440 [Cs]*. Retrieved from http://arxiv.org/abs/1411.0440

Dahroug, A., López-Nores, M., Pazos-Arias, J. J., González-Soutelo, S., Reboreda-Morillo, S. M., & Antoniou, A. (2017). Exploiting relevant dates to promote serendipity and situational curiosity in cultural heritage experiences. In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (pp. 84–88).

Edwards, P., Roberts, I., Green, J., & Lutchmun, S. (2006). Deaths from injury in children and employment status in family: analysis of trends in class specific death rates. *BMJ : British Medical Journal*, *333*, 119.

Erdelez, S. (1997). Information encountering: a conceptual framework for accidental information discovery (pp. 412–421). Presented at the ISIC '96: Proceedings of an international conference on Information seeking in context, Taylor Graham Publishing.

Erdelez, S. (2004). Investigation of information encountering in the controlled research environment. *Information Processing & Management*, *40*, 1013–1025.

Erdelez, S., Heinström, J., Makri, S., Björneborn, L., Beheshti, J., Toms, E., & Agarwal, N. K. (2016). Research perspectives on serendipity and information encountering. *Proceedings of the Association for Information Science and Technology*, *53*, 1–5.

Fazeli, S., Drachsler, H., Bitter-Rijpkema, M., Brouns, F., Vegt, W. van der, & Sloep, P. B. (2017). User-centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg. *IEEE Transactions on Learning Technologies*, *PP*, 1–1.

Foster, A., & Ford, N. (2003). Serendipity and information seeking: an empirical study. *Journal of Documentation*, *59*, 321–340.

Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: evaluating recommender systems by coverage and serendipity (p. 257). ACM Press.

Gries, S. T. (2009). *Quantitative corpus linguistics with R : a practical introduction*. New York, NY : Routledge, 2009. Retrieved from https://www.safaribooksonline.com/library/view/quantitative-corpus-linguistics/9781135895594/Part01.html

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*, 5–53.

Kapetanios, E., Tatar, D., & Sacarea, C. (2013). *Natural Language Processing: Semantic Aspects*. Baton Rouge, UNITED STATES: Chapman and Hall/CRC. Retrieved

from http://ebookcentral.proquest.com/lib/umcolumbia-ebooks/detail.action?docID=1408038

Kefalidou, G., & Sharples, S. (2016). Encouraging serendipity in research: Designing technologies to support connection-making. *International Journal of Human-Computer Studies*, *89*, 1–23.

Krippendorff, K. (2004). *Content analysis : an introduction to its methodology*. Thousand Oaks, Calif.: Sage.

Maccatrozzo, V., van Everdingen, E., Aroyo, L., & Schreiber, G. (2017). Everybody, *More or Less* , likes Serendipity (pp. 29–34). ACM Press.

Makri, S., & Blandford, A. (2012). Coming across information serendipitously – Part 1: A process model. *Journal of Documentation*, *68*, 684–705.

Makri, S., Toms, E. G., McCay-Peet, L., & Blandford, A. (2011). Encouraging Serendipity in Interactive Systems. In *Human-Computer Interaction – INTERACT 2011* (pp. 728–729). Springer, Berlin, Heidelberg.

Mayes, C., & Macleod, C. (1999). When "NAI" means not actually injured. *BMJ : British Medical Journal*, *318*, 1127–1128.

McCay-Peet, L., & Toms, E. (2011). Measuring the dimensions of serendipity in digital environments. *Usos y Gratificaciones: La Medición de Las Dimensiones de La Serendipia En Entornos Digitales.*, *16*, 6–6.

McCay-Peet, L., & Toms, E. (2018). *Researching serendipity in digital information environments*. San Rafael, California: Morgan & Claypool Publishers.

McCay-Peet, L., & Toms, E. G. (2015). Investigating serendipity: How it unfolds and what may influence it. *Journal of the Association for Information Science and Technology*, *66*, 1463–1476.

Niu, X., & Abbas, F. (2017). A Framework for Computational Serendipity. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 360–363). New York, NY, USA: ACM.

Ponampalam, R., Tan, H. H., Ng, K. C., Lee, W. Y., & Tan, S. C. (2009). Demographics of toxic exposures presenting to three public hospital emergency departments in Singapore 2001-2003. *International Journal of Emergency Medicine*, *2*, 25–31.

Rang, M. (1972). The Ulysses syndrome. *Canadian Medical Association Journal*, *106*, 122–123.

Rioux, K. (2005). Information Acquiring-and-Sharing. In *Theories of Information Behavior* (pp. 169–173). Medford, N.J: American Society for Information Science and Technology.

Rubin, V. L., Burkell, J., & Quan-Haase, A. (2011). Facets of serendipity in everyday chance encounters: a grounded theory approach to blog analysis. *Information Research*, *16*, 27–27.

sagacity. (n.d.). *Dictionary.com Unabridged*. Retrieved from http://www.dictionary.com/browse/sagacity

Sun, X., & Sharples, S. (2011). A user-centred mobile diary study approach to understanding serendipity in information research [text]. Retrieved November 22, 2017, from http://www.informationr.net/ir/16-3/paper492.html

Wopereis, I., & Braam, M. (2018). Seeking Serendipity: The Art of Finding the Unsought in Professional Music. In S. Kurbanoğlu, J. Boustany, S. Špiranec, E. Grassian, D. Mizrachi, & L. Roy (Eds.), *Information Literacy in the Workplace* (pp. 503–512). Springer International Publishing.

Wu, W., He, L., & Yang, J. (2012). Evaluating recommender systems. In *Seventh International Conference on Digital Information Management (ICDIM 2012)* (pp. 56–61).

Chapter 5 Conclusion

Serendipity is a unique experience in which the subject encounters information unexpectedly and uses that information in a novel way to achieve valuable results. Because serendipity has been credited with many technological improvements and scientific discoveries, the information science community has deemed it an important information behavior to investigate. This dissertation contributes to the field of information science and the understanding of the information behaviors embedded in the experience of serendipity in several ways. First, this series of studies examines serendipity in a different way than previous studies. While previous studies have elicited participants' reflections on their prior serendipitous experiences or compiled histories, this dissertation uses a content analysis approach to seek evidence of serendipity in existing research documents. Because the documents were written for a different purpose other than the communication of serendipitous events, the language in which the researchers communicate these events is more natural and less prompted than in other, previous studies. Instead of approaching serendipity as a rare occurrence that can only be documented through the collection of anecdotal evidence, we approach the idea of serendipity as a ubiquitous experience that all researchers might encounter. This approach allows us to estimate the prevalence of serendipity in biomedical research as a whole as well as gleaning information from the identified reports.

The series of studies presented in this dissertation contributes to our understanding of serendipity as it is experienced in biomedical research. This series of studies delineates the characterization of serendipity in biomedical research from development of a taxonomy for identifying instances of serendipity through the prediction of relevance of a phrase to serendipity by using of natural language processing tools and qualitative analysis. These studies trace the

progression of understanding as we address the four identified research questions, and endeavor to reach our goal of developing support for serendipity in the research process.

Research Question 1. How do researchers in biomedical and life sciences use terms related to serendipity in their research reporting?

This dissertation explored the semantic meaning communicated by biomedical researchers when using terms that express surprise in their scientific reports. This exploration tested the null hypothesis that all references to surprise in biomedical research publications would be unrelated to the concept of serendipity. The qualitative results of our study refute this hypothesis, as our findings clearly demonstrate that biomedical researchers commonly reveal instances of serendipity in their scientific reports. While some of the references identified contain details of the entire serendipitous experience, from the unexpected encounter through the follow-up, and culminating in valuable outcomes, we also identified some reports that only detail one aspect of a serendipitous experience. If one defines a report of serendipity more narrowly, requiring the encountered information to be pursued, explained, and capitalized upon, the incidence of research articles containing reports of serendipity would be much lower. If we see information encounters as a seed for serendipitous outcomes, only needing exposure to the right conditions to germinate into fruitful outcomes, a much larger population of serendipity reporting is revealed. This raises the question of how the concept of serendipity should be defined by the information sciences community. We should consider the concept of serendipity, not only as it is manifested in traditional information environments but also as it occurs naturally in everyday human experiences. Considering serendipity as it is defined in other academic disciplines, such as art, architecture, and creative practice, could enrich our understanding of serendipity and aid us in developing more robust models of serendipity as an information experience. Further,

defining serendipity more broadly, by examining reported information encounters in addition to fully realized serendipitous discoveries, holds the potential to use these surprising findings to stimulate follow-up by other researchers.

In addition to identifying reports of serendipity in biomedical research, our qualitative analysis also revealed patterns in the researchers relationship to the information encounter described. When the researcher personally encounters the surprising information and follows up on that information, the resulting article is classified as 'research focused'. When the researcher acts on a serendipitous observation made by another researcher, we classify this category of reports as 'inspiration'. When the researcher compiles examples of serendipity in secondary reports and publishes new findings related to patterns they observe in the reports, we classify the instance as 'systematic review'. In all of these cases, additional analysis is done following the information encounter and valuable results are obtained. This is an intriguing finding because current models of serendipity only describe fortuitous outcomes as serendipitous when the outcome is realized by the individual originally involved in the information encounter. Further analysis of the 'inspiration' type of serendipity identified in this study is warranted, and may prove useful in informing the development of systems to support serendipity in scientific endeavors.

Apart from serendipitous encounters that have resulted in a positive outcome, our corpus analysis revealed that researchers routinely report serendipitous experiences immediately following the information encounter. While this finding raises questions regarding why the surprising information was not pursued, it also exposes a large group of surprising observations that might be helpful in populating information recommender systems. Perhaps the key to

turning these initial reports into valuable outcomes lies in our ability to connect those observations to individuals with the sagacity to perform the necessary follow-up.

In addition to the four types of term use that were related to serendipitous events, we also identified seven meanings that were unrelated to serendipity. These include term meanings that convey insignificance or unintendedness, provide historical reference to serendipity, or are used to describe statistical outcomes (i.e. a 'chance' finding). Additional instances that were considered unrelated were negative uses of the surprise descriptor (i.e. 'not surprising'), use of the surprise descriptor in the title of a cited reference, and articles that were returned in the search but did not meet inclusion criteria as research reports. Identification of features of these unrelated meanings could prove useful in developing algorithms for disambiguation of reports of serendipity.

Research Question 2. What proportion of biomedical research publications indicate serendipitous experiences?

Addressing this research question necessitated data that was generalizable, requiring care to be taken in the design of the research methodologies. By analyzing evidence of serendipity in existing documents within a population of research publications, we were able to design sampling procedures that would allow for the generalizability of the quantitative research findings to the population of research articles analyzed. This, in itself, is a novel feature of this study, as the transient nature of serendipitous experiences makes the use of more rigorous experimental approaches challenging.

In exploring the semantic meaning communicated with terms related to the idea of surprise, this dissertation tested the null hypothesis that the use of terms related to surprise are not indicative of reports of serendipity. Our results refute this hypothesis, and reveal that

serendipity is a fairly common information experience for biomedical researchers. Surprise descriptor terms were identified in the full-text of 552,288 of the 4.3 million biomedical research articles indexed by PubMed Central. After analyzing the semantic meaning conveyed by the surprise descriptor terms in our sample, we discovered that 41% of articles containing a surprise descriptor term communicate reports of serendipity. When these findings are extrapolated to the population of 552,288 biomedical research articles indexed by PubMed Central containing the surprise descriptor terms, we estimate nearly 225,000 articles, or approximately 7.5% of all articles indexed by PubMed Central, contain a reference to serendipity. While this finding is fairly impressive, it relies on the definition of serendipity that includes information encounters that have not yet been explored. If we restrict our definition to only those reports with valuable outcomes, the incidence drops to about 75,000 articles, or less than 2% of the articles indexed by PubMed Central. Furthermore, while our findings can be generalized to the population of articles indexed by PubMed Central, they cannot be extended to those fields not indexed by PubMed Central or to academic disciplines outside of biomedical sciences. Additionally, it is unknown whether these findings are applicable to individual academic disciplines indexed within PubMed Central. Further research to explore serendipity reporting by field of study is warranted. Overall, this finding is promising because it demonstrates that biomedical researchers routinely encounter information serendipitously during the course of their scientific investigations. Moreover, these findings support the idea that serendipity, as an information behavior, occurs with sufficient frequency to justify its technological support.

Research Question 3. What syntactic and semantic features distinguish references to serendipitous experiences?

This dissertation explored the syntactic and semantic features exhibited in the scholarly communication of biomedical researchers when reporting serendipitous experiences. Our qualitative analysis focused on identifying features in the sentence and paragraph containing the report of serendipity. We found that reports of serendipity exhibit some consistent syntactic and semantic characteristics, including the use of a modifier that expresses the surprise aspects of the information encounter, the presence of a noun that refers to the type of information that was gleaned from the encounter, and the location of the surprise descriptor within the document. While these initial variables were identified, we have no reason to believe that this is the extent of features that distinguish reports of serendipity from other discussions in biomedical literature. Search for additional features is essential to effectively identifying reports of serendipity in research literature. Furthermore, the features identified are exclusively applicable to reports written in the English language. Additional study of serendipity reporting in other languages would enrich our understanding of how language impacts the understanding of serendipity.

Research Question 4. What features of author reporting are most useful for developing algorithms to automatically compile references to serendipitous experiences?

This dissertation further built on the quantitative findings addressed in research question two by analyzing the syntactic and semantic features exhibited in researcher's narratives for their predictive ability in identifying reports of serendipity. In determining the usefulness of identified semantic and syntactic features in predicting reports of serendipity, this dissertation tested the null hypothesis that coefficients for identified variables would be statistically equivalent to zero, having no more predictive power than random words in the article. We identified three categorical independent variables for analysis: surprise descriptor term used, location of that term within the journal article, and collocation with an information term. Initial bivariate analysis

demonstrated significant relationships between actual serendipitous events and each of the independent variables analyzed. However, the logistic regression results indicated questionable overall model fit for each of the independent variables, despite being statistically reliable predictors. It is possible that the breadth of disciplines included in our sample, despite being identified as biomedical literature, differ significantly in the terminology used to report serendipity. This could result in certain terms being useful predictors in one discipline and not in another, effectively hiding the predictive relevance of a term for one field by its irrelevance in other fields. Additional studies focusing on characterizing reports of serendipity in individual disciplines is needed to determine if domain conventions impact the predictability of serendipity reporting. Furthermore, the predictive value of the variables may be linked to the type of serendipity reported. For example, the predictive value for surprise terms located in the abstract may be higher for 'research focused'- type serendipity reports, whereas location in the introduction or literature review may have higher predictability for 'inspiration'- type serendipity reports. Further study is needed to better delineate predictive factors as they relate to the type of serendipity reported.

Overall, as a screening tool for the identification of reports of serendipity, full-text search for the surprise descriptors identified in this study exhibited a high sensitivity in that it reliably captures references to serendipity. Unfortunately, as a predictive model, surprise descriptors exhibit poor specificity, returning a large proportion of articles that do not report serendipity in addition to the capturing valid instances, as only 41% of the returned results were actually referring to a serendipitous experience. The additional predictors of collocation with an information term and surprise detector location within the document improved the specificity of the model, but greatly decreased its sensitivity. Additional predictive features are needed to make

the automatic compilation of reports of serendipity feasible as a support mechanism for serendipity in information recommender systems.

Limitations

While care was taken when compiling the corpus to insure that the articles sampled were representative of the full depth and breadth of articles indexed by PubMed Central, our findings are only applicable to those article indexed by PubMed Central and may not be representative of fields outside of the biomedical sciences or of academic literature not indexed by PubMed Central. Additionally, the breadth of fields indexed by PubMed Central and our attempts to fully represent that population may have inadvertently hidden the effectiveness of the identified predictive variables, as particular features may be representative of some disciplines indexed in PubMed Central and not others. Finally, as a corpus study, this research reveals patterns evident in English-language reported biomedical research. Such patterns may not hold true for publications in other languages

Key Findings

Overall, this line of research is crucial in advancing our understanding of serendipity as an information behavior and in the development of technological systems to support and promote serendipitous information encounters. Our research has demonstrated that:

1. Serendipitous experiences occur much more frequently in the course of conducting biomedical research than previously expected, with an estimate of nearly 225,000 articles, or approximately 7.5% of the articles indexed by PubMed Central, containing a reference to serendipity.
2. Reports of serendipity in biomedical sciences are not limited to those information experiences where the researcher identifies unexpected information, pursues an

understanding of that information and builds on that understanding to achieve a valuable outcome. Serendipity is reported when the researcher is inspired by an unusual finding reported by another researcher, when the researcher is analyzing the impact of serendipity on a facet of their discipline, and when the researcher has identified an anomaly in their findings, but has not yet acted on that information.

3. Reports of serendipity in the biomedical literature exhibit some consistent linguistic characteristics, including the use of a modifier that expresses the surprise aspects of the information encounter, the presence of a noun that refers to the type of information that was gleaned from the encounter, and the location of the surprise descriptor within the article, that hold promise for automatic compilation.
4. Full-text search for the identified surprise descriptors exhibits a high sensitivity in capturing reports of serendipity, but additional predictors are needed to remove human analysis of meaning from the identification process.

Bibliography

Allen, C., Erdelez, S., & Marinov, M. (2013). Looking for Opportunistic Discovery of Information in Recent Biomedical Research – A Content Analysis. In *Proceedings of the 76th ASIS&T Annual Meeting* (Vol. 49). Montreal: Richard B. Hill.

Allen, C. M., & Erdelez, S. (2018). Distraction to Illumination: Mining Biomedical Publications for Serendipity in Research. *Proceedings of the American Society for Information Science and Technology*, *55*, 11–19.

Badgaiyan, R. D., & Wack, D. (2011). Evidence of dopaminergic processing of executive inhibition. *PloS One*, *6*(12), e28075.

Barnett, A. H. (2011). Diabetes-science, serendipity and common sense. *Diabetic Medicine*, *28*(11), 1289–1299.

Barreiro, C., Martín, J. F., & García-Estrada, C. (2012). Proteomics shows new faces for the old penicillin producer Penicillium chrysogenum. *Journal of Biomedicine & Biotechnology*, *2012*, 105109.

Bates, M. J. (2005). Berrypicking. In *Theories of Information Behavior* (pp. 69–74). Medford, N.J: American Society for Information Science and Technology.

Buckland, M. K. (2017). *Information and society*. Cambridge, Massachusetts: The MIT Press.

Buechel, H. M., Popovic, J., Searcy, J. L., Porter, N. M., Thibault, O., & Blalock, E. M. (2011). Deep sleep and parietal cortex gene expression changes are related to cognitive deficits with age. *PloS One*, *6*(4), e18387.

Campanario, J. M. (1996). UsingCitation Classics to study the incidence of serendipity in scientific discovery. *Scientometrics*, *37*(1), 3–24.

Cannon, W. B. (1965). Gains from Serendipity. In *The way of an investigator : a scientist's experiences in medical research* (pp. 68–78). New York: Hafner.

Case, D. O. (2012). *Looking for Information : a survey of research on information seeking, needs and behavior*. Bingley, UK: Emerald Group Pub.

Chang, S.-J. (2005). Chang's Browsing. In *Theories of Information Behavior* (pp. 69–74). Medford, N.J: American Society for Information Science and Technology.

Corneli, J., Jordanous, A., Guckelsberger, C., Pease, A., & Colton, S. (2014). Modelling serendipity in a computational context. *ArXiv:1411.0440 [Cs]*.

Creswell, J. W. (2014). *Research design: qualitative, quantitative, and mixed methods approaches* (4th ed). Thousand Oaks: SAGE Publications.

Dahroug, A., López-Nores, M., Pazos-Arias, J. J., González-Soutelo, S., Reboreda-Morillo, S. M., & Antoniou, A. (2017). Exploiting relevant dates to promote serendipity and situational curiosity in cultural heritage experiences. In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (pp. 84–88).

Davis, R. A., Wetzel, N., & Davis, L. (1956). An Analysis of the Results of Treatment of Intracranial Vascular Lesions by Carotid Artery Ligation. *Annals of Surgery*, *143*(5), 641–648.

Edwards, P., Roberts, I., Green, J., & Lutchmun, S. (2006). Deaths from injury in children and employment status in family: analysis of trends in class specific death rates. *BMJ : British Medical Journal*, *333*(7559), 119.

Ellis, D. (1989). A behavioural model for information retrieval system design. *Journal of Information Science*, *15*(4–5), 237–247.

Erdelez, S., Basic, J., & Levitov, D. D. (2011). Potential for inclusion of information encountering within information literacy models. *Information Research*, *16*(3).

Erdelez, Sanda. (1997). Information encountering: a conceptual framework for accidental information discovery (pp. 412–421). Presented at the ISIC '96: Proceedings of an international conference on Information seeking in context, Taylor Graham Publishing.

Erdelez, Sanda. (1999). Information Encountering: It's More Than Just Bumping into Information. *Bulletin of the American Society for Information Science and Technology*, *25*(3), 26–29.

Erdelez, Sanda. (2004). Investigation of information encountering in the controlled research environment. *Information Processing & Management*, *40*(6), 1013–1025.

Erdelez, Sanda. (2009). Information Encountering. In *Theories of Information Behavior* (pp. 179–184). Medford, N.J.: Information Today Inc.

Erdelez, Sanda, & Basic, J. (2011). Potential for inclusion of information encountering within information literacy models. *Information Research*, *16*(3).

Erdelez, Sanda, Basic, J., & Levitov, D. D. (2011). Potential for inclusion of information encountering within information literacy models. *Information Research*, *16*(3).

Erdelez, Sanda, Heinström, J., Makri, S., Björneborn, L., Beheshti, J., Toms, E., & Agarwal, N. K. (2016). Research perspectives on serendipity and information

encountering. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–5.

Erdelez, Sanda, Marinov, M., & Allen, C. (2012). Knowledge Discovery System for Research Hypothesis Generation from Serendipitous Findings: A Feasibility Study. In *Proceedings - AMIA 2012 Annual Symposium*. Chicago: American Medical Informatics Association.

Fazeli, S., Drachsler, H., Bitter-Rijpkema, M., Brouns, F., Vegt, W. van der, & Sloep, P. B. (2017). User-centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg. *IEEE Transactions on Learning Technologies*, *PP*(99), 1–1.

Fisher, K. E. (2005). Information Grounds. In *Theories of Information Behavior* (pp. 185–190). Medford, N.J: American Society for Information Science and Technology.

Fisher, K. E., Erdelez, S., & McKechnie, L. (2005). *Theories of information behavior*. Medford, N.J.: Published for the American Society for Information Science and Technology by Information Today.

Foster, A., & Ellis, D. (2014). Serendipity and its study. *Journal of Documentation; Bradford*, *70*(6), 1015–1038.

Foster, A., & Ford, N. (2003). Serendipity and information seeking: an empirical study. *Journal of Documentation*, *59*(3), 321–340.

Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: evaluating recommender systems by coverage and serendipity (p. 257). ACM Press.

Giesy, J., Newsted, J., & Oris, J. (2013). Photo-enhanced toxicity: serendipity of a prepared mind and flexible program management. *Environmental Toxicology and Chemistry*, *32*(5), 969–971.

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*, R86.

Gries, S. T. (2009). *Quantitative corpus linguistics with R : a practical introduction*. New York, NY : Routledge, 2009.

Hargrave-Thomas, E., Yu, B., & Reynisson, J. (2012). Serendipity in anticancer drug discovery. *World Journal of Clinical Oncology*, *3*(1), 1–6.

Harris, E. L. (1995). *The shadowmakers: A history of radiologic technology* (1st edition). Albuquerque, N.M: American Society of Radiologic Technologists.

Hemphill, R. E., & Stengel, E. (1940). A Study on Pure Word-Deafness. *Journal of Neurology and Psychiatry*, *3*(3), 251–262.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53.

Houge, G., & Hennekam, R. C. M. (2009). Reply to Happle. *European Journal of Human Genetics*, *17*(7), 882.

Iaquinta, L., Gemmis, M. d, Lops, P., Semeraro, G., Filannino, M., & Molino, P. (2008). Introducing Serendipity in a Content-Based Recommender System. In *2008 Eighth International Conference on Hybrid Intelligent Systems* (pp. 168–173).

Kapetanios, E., Tatar, D., & Sacarea, C. (2013). *Natural Language Processing: Semantic Aspects*. Baton Rouge, UNITED STATES: Chapman and Hall/CRC.

Kefalidou, G., & Sharples, S. (2016). Encouraging serendipity in research: Designing technologies to support connection-making. *International Journal of Human-Computer Studies*, *89*, 1–23.

Khalili, A., van Andel, P., van den Besselaar, P., & de Graaf, K. A. (2017). Fostering Serendipitous Knowledge Discovery Using an Adaptive Multigraph-based Faceted Browser. In *Proceedings of the Knowledge Capture Conference* (pp. 15:1–15:4). New York, NY, USA: ACM.

Krippendorff, K. (2004). *Content analysis : an introduction to its methodology*. Thousand Oaks, Calif.: Sage.

Kumpulainen, S. W., & Kautonen, H. (2017). Accidentally Successful Searching: Users' Perceptions of a Digital Library (pp. 257–260). ACM Press.

Leckie, G. J., Pettigrew, K. E., & Sylvain, C. (1996). Modeling the Information Seeking of Professionals: A General Model Derived from Research on Engineers, Health Care Professionals, and Lawyers. *The Library Quarterly: Information, Community, Policy*, *66*(2), 161–193.

Lee, Y.-C. (2011). *Serendipity in scientific discoveries: Some examples in glycosciences* (Vol. 705).

Lesko, L. (2017). Efficacy From Strange Sources. *Clinical Pharmacology & Therapeutics*, *103*(2), 253–261.

Ligon, B. L. (2004). Penicillin: Its Discovery and Early Development. *Seminars in Pediatric Infectious Diseases*, *15*(1), 52–57.

Maccatrozzo, V., van Everdingen, E., Aroyo, L., & Schreiber, G. (2017). Everybody, *More or Less* , likes Serendipity (pp. 29–34). ACM Press.

Makri, S., & Blandford, A. (2012). Coming across information serendipitously – Part 1: A process model. *Journal of Documentation*, *68*(5), 684–705.

Makri, S., Toms, E. G., McCay-Peet, L., & Blandford, A. (2011). Encouraging Serendipity in Interactive Systems. In *Human-Computer Interaction – INTERACT 2011* (pp. 728–729). Springer, Berlin, Heidelberg.

Marton, F. (1986). Phenomenography—A Research Approach to Investigating Different Understandings of Reality. *Journal of Thought*, *21*(3), 28–49.

Mayes, C., & Macleod, C. (1999). When "NAI" means not actually injured. *BMJ : British Medical Journal*, *318*(7191), 1127–1128.

Mayor, S. (2010). Serendipity and cell biology. *Molecular Biology of the Cell*, *21*(22), 3807–3808.

McBirnie, A. (2008). Seeking serendipity: the paradox of control. *Aslib Proceedings*, *60*(6), 600–618. https://doi.org/10.1108/00012530810924294

McCay-Peet, L., & Toms, E. (2011). Measuring the dimensions of serendipity in digital environments. *Usos y Gratificaciones: La Medición de Las Dimensiones de La Serendipia En Entornos Digitales.*, *16*(3), 6–6.

McCay-Peet, L., & Toms, E. (2015). Investigating serendipity: How it unfolds and what may influence it. *Journal of the Association for Information Science and Technology*, *66*(7), 1463–1476.

McCay-Peet, L., & Toms, E. (2017). Researching Serendipity in Digital Information Environments. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *9*(6), i–91. https://doi.org/10.2200/S00790ED1V01Y201707ICR059

McCay-Peet, L., & Toms, E. (2018). *Researching serendipity in digital information environments*. San Rafael, California: Morgan & Claypool Publishers.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (pp. 1097–1101). New York, NY, USA: ACM.

Merton, R. K., & Barber, E. G. (2004). *The travels and adventures of serendipity: a study in historical semantics and the sociology of science*. Princeton, N.J: Princeton University Press.

Meyer-Rochow, V. B. (2009). Food taboos: their origins and purposes. *Journal of Ethnobiology and Ethnomedicine*, *5*, 18.

Million, A. J., O'Hare, S., Lowrance, N., & Erdelez, S. (2013). Opportunistic discovery of information and millennials: An exploratory survey. *Proceedings of the ASIST Annual Meeting*, *50*(1).

Miwa, M., Egusa, Y., Saito, H., Takaku, M., Terai, H., & Kando, N. (2011). A method to capture information encountering embedded in exploratory web searches. *Information Research*, *16*(3).

Mould, R. F. (1995). Invited review: Rontgen and the discovery of X-rays. *British Journal of Radiology*, *68*(815), 1145–1176.

Myers, J. D., Allison, T. C., Bittner, S., Didier, B., Frenklach, M., Green, W. H., … Yang, C. (2004). A collaborative informatics infrastructure for multi-scale science. In *Proceedings of the Second International Workshop on Challenges of Large Applications in Distributed Environments, 2004. CLADE 2004.* (pp. 24–33). https://doi.org/10.1109/CLADE.2004.1309089

Naumer, C. (2005). Flow Theory. In *Theories of Information Behavior* (pp. 153–157). Medford, N.J: American Society for Information Science and Technology.

Niu, X., & Abbas, F. (2017). A Framework for Computational Serendipity. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 360–363). New York, NY, USA: ACM.

Oshmyansky, A. R., Mahammedi, A., Dackiw, A., Ball, D. W., Schulick, R. D., Zeiger, M. A., & Siegelman, S. S. (2013). Serendipity in the Diagnosis of Pheochromocytoma. *Journal of Computer Assisted Tomography*, *37*(5), 820.

Pálsdóttir, Á. (2010). The connection between purposive information seeking and information encountering: A study of Icelanders' health and lifestyle information seeking. *Journal of Documentation*, *66*(2), 224–244.

Panahi, S., Watson, J., & Partridge, H. (2016). Information encountering on social media and tacit knowledge sharing. *Journal of Information Science*, *42*(4), 539–550.

Patten, M. L. (2005). *Understanding research methods: an overview of the essentials*. Glendale, Calif.: Pyrczak Pub.

Patton, M. Q. (2001). *Qualitative Research & Evaluation Methods* (3rd ed.). Sage Publications, Inc.

Ponampalam, R., Tan, H. H., Ng, K. C., Lee, W. Y., & Tan, S. C. (2009). Demographics of toxic exposures presenting to three public hospital emergency departments in Singapore 2001-2003. *International Journal of Emergency Medicine*, *2*(1), 25–31.

Princeton University. (2010). WordNet | A Lexical Database for English. Retrieved August 27, 2018, from https://wordnet.princeton.edu/

Raja, P. R. (2016). Cyanoacrylate Adhesives: A Critical Review. *Reviews of Adhesion and Adhesives*, *4*(4), 398–416.

Rang, M. (1972). The Ulysses syndrome. *Canadian Medical Association Journal*, *106*(2), 122–123.

Raskin, J. D. (2012). Evolutionary constructivism and humanistic psychology. *Journal of Theoretical and Philosophical Psychology*, *32*(2), 119–133.

Rioux, K. (2005). Information Acquiring-and-Sharing. In *Theories of Information Behavior* (pp. 169–173). Medford, N.J: American Society for Information Science and Technology.

Rowley, J. (1998). What is information? *Information Services & Use*, *18*(4), 243.

Rubanyi, G. M. (2011). The discovery of endothelin: The power of bioassay and the role of serendipity in the discovery of endothelium-derived vasocative substances. *Pharmacological Research*, *63*(6), 448–454.

Ruben, B. D. (1992). The communication-information relationship in system-theoretic perspective. *Journal of the American Society for Information Science*, *43*(1), 15–27.

Rubin, V. L., Burkell, J., & Quan-Haase, A. (2011). Facets of serendipity in everyday chance encounters: a grounded theory approach to blog analysis. *Information Research*, *16*(3), 27–27.

sagacity. (n.d.). *Dictionary.com Unabridged*. Retrieved from http://www.dictionary.com/browse/sagacity

Schwager, E. (2000). Little-known facts about little-known people. *Drug News and Perspectives*, *13*(2), 126–128.

Serenko, A., & Dumay, J. (2017). Citation classics published in knowledge management journals. Part III: author survey. *Journal of Knowledge Management*, *21*(2), 330–354.

Shapiro, G. (1986). *A skeleton in the darkroom: stories of serendipity in science* (1st ed). San Francisco: Harper & Row.

Siris, S. G. (2011). Searching For Serendipity. *The Journal of Clinical Psychiatry*, *72*(8), 1156–1157.

Smith, C. E. (2011). Geospatial encountering: Opportunistic information discovery in web-based GIS environments. *Proceedings of the ASIST Annual Meeting*, *48*.

Soldati, G., Smargiassi, A., Mariani, A. A., & Inchingolo, R. (2017). Novel aspects in diagnostic approach to respiratory patients: is it the time for a new semiotics? *Multidisciplinary Respiratory Medicine*, *12*, 15.

Solomon, Y., & Bronstein. (2015). Serendipity in legal information seeking behavior: Chance encounters of family-law advocates with court rulings. *Aslib Journal of Information Management*, *68*(1), 112–134.

Sun, X., Sharples, S., & Makri, S. (2011). A user-centred mobile diary study approach to understanding serendipity in information research. *INFORMATION RESEARCH-AN INTERNATIONAL ELECTRONIC JOURNAL*, *16*(3).

Sundarakumar, D. K. (2011). Non-destructive testing in DNB/MD examination. *The Indian Journal of Radiology & Imaging*, *21*(3), 239–240.

Talip, B. A. (2015). IT professionals' information behaviour on twitter. *Libres*, *25*(2), 86–102.

Tluczek, A., Clark, R., McKechnie, A. C., Orland, K. M., & Brown, R. L. (2010). Task-Oriented and Bottle Feeding Adversely Affect the Quality of Mother-Infant Interactions Following Abnormal Newborn Screens. *Journal of Developmental and Behavioral Pediatrics : JDBP*, *31*(5), 414–426.

U.S. National Library of Medicine. (2018, June 19). PMC FAQs. Retrieved August 27, 2018, from https://www.ncbi.nlm.nih.gov/pmc/about/faq/

Van Andel, P. (1994). Anatomy of the Unsought Finding. Serendipity: Origin, History, Domains, Traditions, Appearances, Patterns and Programmability. *The British Journal for the Philosophy of Science*, *45*(2), 631–648.

Wang, X., Erdelez, S., Allen, C., Anderson, B., Cao, H., & Shyu, C.-R. (2011). Medical image describing behavior: A comparison between an expert and novice. In *ACM International Conference Proceeding Series* (pp. 792–793). Seattle, WA.

Wenger, E. (2000). Communities of Practice and Social Learning Systems. *Organization*, *7*(2), 225–246.

Wilkinson, C., & Weitkamp, E. (2013). A Case Study in Serendipity: Environmental Researchers Use of Traditional and Social Media for Dissemination. *PLOS ONE*, *8*(12), e84339.

Williamson, K. (1998). Discovered by chance: The role of incidental information acquisition in an ecological model of information use. *Library & Information Science Research*, *20*(1), 23–40.

Williamson, K. (2005). Ecological Theory of Human Information Behavior. In *Theories of Information Behavior* (pp. 128–137). Medford, N.J: American Society for Information Science and Technology.

Wilson, T. D. (2000). Human information behavior. *Informing Science*, *3*(2), 49–55.

Wong, C. C. Y., Caspi, A., Williams, B., Houts, R., Craig, I. W., & Mill, J. (2011). A Longitudinal Twin Study of Skewed X Chromosome-Inactivation. *PLoS ONE*, *6*(3).

Wopereis, I., & Braam, M. (2018). Seeking Serendipity: The Art of Finding the Unsought in Professional Music. In S. Kurbanoğlu, J. Boustany, S. Špiranec, E. Grassian, D. Mizrachi, & L. Roy (Eds.), *Information Literacy in the Workplace* (pp. 503–512). Springer International Publishing.

Wu, W., He, L., & Yang, J. (2012). Evaluating recommender systems. In *Seventh International Conference on Digital Information Management (ICDIM 2012)* (pp. 56–61).

Yadamsuren, B., & Erdelez, S. (2010). Incidental exposure to online news. *Proceedings of the ASIST Annual Meeting*, *47*.

Yadamsuren, B. (2013). Potential of inducing serendipitous news discovery in social gaming environment. *Proceedings of the ASIST Annual Meeting*, *50*(1).

Yadamsuren, B., & Erdelez, S. (2016). Incidental Exposure to Online News. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *8*(5), i–73.

Young, G., Ashdown, D., Arnold, A., & Subramonian, K. (2008). Serendipity in urology. *BJU International*, *101*(4), 415–416.

Vita

Carla McCaghren Allen graduated from the University of Missouri with a Bachelor of Health Science degree in Radiologic Sciences in 1993. She earned ARRT credentials in Radiography and Computed Tomography, practicing medical imaging at the Truman VA Hospital in Columbia, Missouri until 1999. At that time, Mrs. Allen transitioned into Radiography education, serving as Clinical Education Coordinator at Nichols Career Center and Program Director at State Fair Community College before joining the faculty of the University of Missouri Radiography Program in 2006. She earned the Master of Education degree in Career and Technical Education from the University of Missouri in 2002.

Mrs. Allen currently serves as an Associate Clinical Professor of Radiography in the School of Health Professions at the University of Missouri. . Her research focuses on human information behavior, with a particular interest in information encountering, serendipity and image-based information behaviors. Her research productivity was recently recognized with the Kristofer and Lori Hagglund Early Career Award.