Humboldt-Universität zu Berlin

Institut für Bibliotheks- und Informationswissenschaft

DISSERTATION

BEYOND THE PAYWALL:

A Multi-Sited Ethnographic Examination of the Information-Related Behaviors of Six Scientists

zur Erlangung des akademischen Grades

Doctor philosophiae (Dr. phil.)

Philosophische Fakultät I

Stephanie Krueger

Der Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz
Die Dekanin der Philosophischen Fakultät I
Prof. Dr. Gabriele Metzler

Gutachter: 1. Prof. Michael Seadle, PhD

2. Prof. Dr. Jörg Niewöhner

Datum der Einreichung: 06/05/2016 Datum der Disputation: 22/07/2016

Erklärung über die selbstständige Abfassung meiner Dissertation

Hiermit erkläre ich, STEPHANIE KRUEGER, Matrikel-Nr: 556513

dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.
Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.
Die Dissertation wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt oder veröffentlicht.

Prag, den 6. Mai 2016

Stephanie Krueger

Zusammenfassung: Beyond the Paywall: A Multi-Sited Ethnographic Examination of the Information-Related Behaviors of Six Scientists

In dieser Dissertation untersuche ich die Forschungswege von sechs Wissenschaftlern, die in verschiedenen Disziplinen und Institutionen in den Vereinigten Staaten und in der Tschechischen Republik arbeiten. Um dies zu tun, verwende ich sogenannte *„multi-sited“* ethnographisch-methodische Strategien (d.h. Strategien, die Anthropologen verwenden, um Kulturen an zwei oder mehr geografischen Standorten zu vergleichen), mit dem Ziel, informationsbezogene Verhaltensweisen dieser Wissenschaftler im *global vernetzten akademischen Umfeld* zu untersuchen, englisch abgekürzt „GNAE“, ein Begriff, der sich speziell auf die komplexe Bricolage von Netzwerkinfrastrukturen, Online-Informationsressourcen und Tools bezieht, die Wissenschaftler heutzutage nutzen, d.h. die weltweite akademische e-IS, oder akademische Infrastruktur (Edwards et al. 2013).

Die zentrale Forschungsfrage (RQ1), die in dieser Dissertation beantwortet wird, ist: Gibt es, gemäß der *multi-sited* ethnographischen Analyse der beteiligten Wissenschaftler in dieser Studie—Personen, die Forschung in verschiedenen Disziplinen und Institutionen sowie an unterschiedlichen Standorten betreiben—Hinweise darauf, dass ein signifikanter Anteil der nicht-institutionellen/informellen informationsbezogenen Forschung über Mechanismen im GNAE, die nicht von Bibliotheken unterstützt werden, betrieben wird, sowie (RQ2): Was für Muster sind vorhanden und wie beziehen sie sich auf informationswissenschaftliche und andere sozialwissenschaftliche Theorien? Und drittens (RQ3): Haben die Resultate praxisnahe Bedeutungen für die Entwicklung von Dienstleistungen in wissenschaftlichen Bibliotheken?

Ethnographische Strategien sind bisher noch nicht in der Informationswissenschaft (IS) eingesetzt worden, um Fragen dieser Art zu untersuchen. Die Ergebnisse zeigen, dass eine informelle Informationsexploration nur bei zwei Wissenschaftlern, die mit offenen Daten und Tools einer verteilten Computing-Infrastruktur arbeiten, zu finden ist.

Stichworte: multi-sited ethnography, information behavior, global vernetztes akademisches Umfeld, Chemieinformatik

Abstract: **Beyond the Paywall: A Multi-Sited Ethnographic Examination of the Information-Related Behaviors of Six Scientists**

In this dissertation I examine the pathways of information exploration and discovery of six scientists working in different research disciplines affiliated with several academic institutions in the United States and in the Czech Republic. To do so, I utilize multi-sited ethnographic methodological strategies (i.e., strategies developed by anthropologists to compare cultures across two or more geographic locations) to examine the information-related behaviors of these scholars within the global networked academic environment (GNAE), a term which specifically refers to the complex bricolage of network infrastructures, online information resources, and tools scholars use to perform their research today (i.e., the worldwide academic e-IS, or academic infrastructure [Edwards et al. 2013]).

The central research question (RQ1) to be answered in this dissertation: *According to the multi-sited ethnographic analysis of scientists participating in this study—individuals conducting research in various disciplines at different institutions in several geographical locations—is there evidence indicating a significant allotment of non-institutional/informal information-related exploration and discovery occurring beyond official library-supported mechanisms in the GNAE?*, and—part two (RQ2) of the central research question—*What (if any) patterns are exhibited and how do these patterns relate to information science (IS) and other social science theories?* Both RQ1 and RQ2 are exploratory. I additionally ask (RQ3): *What might all this mean in the applied sense?* by showing examples of services piloted during the research process in response to my observations in the field.

Multi-sited ethnographic strategies have not yet been employed in IS, as of the date of publication of this thesis, to examine such questions. Results indicate informal information exploration occurring only with two scientists who use of open data and tools on a distributed computing infrastructure.

Keywords: multi-sited ethnography, information behavior, global networked academic environment, cheminformatics

Acknowledgements

To my research participants

To my colleagues at the National Library of Technology in Prague

To my mentor, Prof. Michael Seadle

Many thanks to my second reader, Prof. Jörg Niewöhner

Inspired by:

- G. Flaubert’s *Bouvard et Pécuchet*
- K. Blum’s *Alltag in Amerika*
- Milan and Jan, who introduced me to the Fast Fourier Transform and concrete failure
- The music of SCH
- The encyclopedic superficiality of Jarda and Franta
- Sasha’s confabulations

Table of Contents

List of Tables

List of Figures

Abbreviations

CPU: Central processing unit

CSCW: Computer-Supported Cooperative Work

CTU: Czech Technical University in Prague

GNAE: Global Networked Academic Environment

e-IS: e-Infrastructures

FB: Facebook

Google CZ: Google search engine, Czech version

Google US: Google search engine, US version

GS: Google Scholar

IDE: Integrated Development Environment

IEEE: Institute of Electrical and Electronics Engineers

IP address: Internet Protocol Address

IRB: Institutional Review Board

IS: Information Science

KBT: Knowledge-Based Trust algorithm (Google)

KV: Knowledge Vault (Google)

LIS: Library and Information Science

Master’s L: Carnegie Basic Classification Master's Colleges and Universities (larger programs)

MMR: Mixed Method Research

NIS: US National Institutes of Health

NTK: Czech National Library of Technology

OChem: Online Chemical Database with modeling environment

OSS: Open source software

PLOS: Public Library of Science

PNAS: Proceedings of the National Academy of Sciences

PubChem: US National Library of Medicine, National Center for Biotechnology Information,

PubMed: US National Library of Medicine, National Institutes of Health

RDS: Respondent-driven sampling

RU/H: Carnegie Basic Classification Research Universities (high research activity),

RU/VH: Carnegie Basic Classification Research University (very high research activity), SEO: Search engine optimization

SS: Snowball sampling

SS/RDS: Snowball sampling/respondent driven sampling

STEM: Science, Technology, Engineering, and Mathematics

STS: Science and Technology Studies

VPN: Virtual Private Network

WOS: Web of Science

Chapter 1. Introduction

1.1 Origins of Research

Anyone working in the information science (IS) professions over the past two decades—be it as a librarian, scholar, resource provider, or system designer—can attest to the significant environmental and infrastructural changes witnessed in various manifestations since the late 1990s, with information retrieval (IR) systems reaching a stage of global interconnectedness and human information behavior research manifesting itself in various forms (Saracevic 2010). This has meant that information science itself has gone through an extended period of examination of and debate about its theoretical and disciplinary underpinnings, with no unifying paradigms or sets of theories emerging to fully frame IS research agendas—a recent example of this is illustrated in Bawden et al.'s review of quantum IS (Bawden et al. 2015). Cibangu (2013) describes the overall situation in detail. For IS researchers, this means one must select a path through a dense and sometimes confusing mélange of theories and research methods, both quantitative and qualitative, in order to address research questions which can cover a wide variety of topics and traditions.

There does, however, appear to be some theoretical consensus around Bates' Three Big Questions for IS, one of which being the *social question: how do people relate to, seek, and use information* (Saracevic 2010, p. 8) for research questions involving the interaction of IR systems with individuals in the context of a global networked academic environment (GNAE), an umbrella term for various names which have been used for this complex bricolage of network infrastructures, online information resources, and tools scholars use to perform their research today. A full definition of the GNAE is provided in the next section of this dissertation. Cronin (2008, p. 466) christens this examination of social questions in IS a "sociological turn in information science" and Seadle (2011) concisely describes where and how qualitative ethnographic research might serve IS in relation to technology and user studies. While specific techniques can vary, IS researchers recognize that individual information behavior does not occur in a black box separate from the environment(s), infrastructure(s), or community(-ies) in which an individual works. However, while we in IS may now know some things about representing and retrieving information—and even perhaps are beginning to understand a little about why different kinds of people work with different kinds of information in the way they do today (Saracevic 2010, p. 11)—as a discipline, there still remains much work to do in order to learn more about what the intersection of IR systems and individuals means within a larger social context; indeed, "[o]ur discipline has been more concerned with the facilitation of communication processes than with their explanation" (Hjørland and Albrechtson 1995, p. 409 citing Boyce and Kraft 1985, p. 165). To remedy this, an increasing number of IS researchers are turning to anthropology for methodological and theoretical assistance (e.g., Emary 2014; Khoo et al. 2012; Wakimoto 2013) in grappling with the symbolic interactions and/or patterns gleaned from observing and describing information behaviors—the interplay between actors, information, and systems (broadly defined to include networks, infrastructures, institutions) in various settings—and positing theories about their meaning(s). I will, in this dissertation, take an anthropological stance to my research questions within IS; in doing so, I will also attempt to avoid the "unfortunate reinvention of social science precepts" which has, according to Sandstorm and Sandstorm (1995, p. 163), plagued some earlier qualitative IS studies.

This dissertation considers overarching questions surrounding the meaning of observed actors, systems, and behaviors by examining how six scientists work with information in multiple locations, in various disciplines, in the context of the GNAE. I utilize multi-sited ethnographic methodological strategies—inherently comparative in nature because they involve investigating multiple populations at multiple sites of observation—to consider if, how, and why the scientists I observed are circumventing traditional providers of information such as libraries as they conduct their research. In other words, I will examine the *information lifeworlds* of the individual scientists as well as the information-related systems and associations in which they work (Marcus 1995). In the process, I attempt to identify patterns of meaning—both local and global—from the data I have gathered, and will explain how, embedded in a science and technology library throughout the formal phase of the research project (2013-2015), I attempted to create appropriate service responses to the patterns of meaning I identified.

All the scientists with whom I have worked since embarking on my study in late 2012 are from disciplines in which I have no formal academic training: my key informants, all formal research participants, are active researchers in the areas of theoretical and experimental physics, bio-/cheminformatics, and immunology/infectious diseases. Other scientists, not formal research participants but representative members of the science and technology communities with whom I have worked since 2012, have varied backgrounds as well, including computer science, chemical engineering, mathematics, and civil engineering. While the majority of participants I studied live in either the United States or the Czech Republic—the former being my country of birth and the latter, where I currently live and work—the daily activities of all participants occurs on an international stage, conducted locally and simultaneously dispersed across institutional, national, and geographical boundaries.

As appropriate for a contemporary ethnography, I will narrate this dissertation in the first person in order to reveal aspects of my thought processes as a form of reflexive self-presentation (Marcus 1995). I myself am an information scientist who found herself, prior to 2012, working within the GNAE since 2001—primarily with teams of software developers and global communities of users (librarians, professors, and students representing a wide range of academic disciplines) of the digital libraries of text, sound, and images initially funded by The Andrew W. Mellon Foundation (JSTOR, DRAM, and Artstor, respectively). Because this professional work between 2001 and 2011 was quite pragmatic in nature and included developing and implementing specific solutions for the aforementioned organizations, I had little time to formally contemplate the broader implications of what the rapid development and adaption of such digital academic tools meant to our users—and even to myself. But when the opportunity arose to take a more research-oriented position at a library in the heart of a science and engineering campus in Central Europe, I decided the time had come for me to systematically learn more about the global community of academics I had served with for more than a decade. In some ways, therefore, the origins of this dissertation date back almost fifteen years—I have been embedded, so to speak, in the GNAE in various ways since the turn of this century. And since 2012, I have been an acute observer of scientists at various stages in their careers.

1.2 Problem Statement

In this dissertation I examine the pathways of information exploration and discovery of scientists working in different research disciplines affiliated with several academic institutions in the United States and in the Czech Republic. To do so, I utilize multi-sited ethnographic methodological strategies to examine the information-related behaviors of these scholars within the GNAE, a term which reflects the ontological status of information in a global network rather than the pre-network information systems and structures which existed prior the Internet. The term GNAE encompasses previous descriptions of the networked environment referred to since the advent of the Internet with various terms ranging from *fluid hypersphere* (Tsekeris 2008) to *dataspace* to *e-Infrastructures (e-IS)* and beyond (see notably Borgman et al. 2012; Pollock and Williams 2010; Skenderija 2008; Willcocks and Whitley 2009). The GNAE specifically refers to the complex bricolage of network infrastructures, online information resources, and tools scholars use to perform their research today (i.e., the worldwide academic e-IS, or academic infrastructure (Edwards et al. 2013).

The central research question (RQ1) to be answered in this dissertation: *According to the multi-sited ethnographic analysis of scientists participating in this study—individuals conducting research in various disciplines at different institutions in several geographical locations—is there evidence indicating a significant allotment of non-institutional/informal information-related exploration and discovery occurring beyond official library-supported mechanisms in the GNAE*, and—part two (RQ2) of the central research question—*What (if any) patterns are exhibited and how do these patterns relate to IS and other social science theories*? Both RQ1 and RQ2 are exploratory. I additionally ask (RQ3): *What might all this mean in the applied sense?* by showing examples of services piloted during the research process in response to my observations in the field.

The concept of non-institutional, information-related exploration and discovery encompasses the utilization of all possible recorded outcomes of scholarly communication transmitted within GNAE—scholarly publications, repositories, drafts, course materials, demonstrations and presentations, correspondence, research notes, chats, blogs, posts, online forums, datasets and data visualizations (interactive 3D models and simulations), software, multimedia recordings, and others—regardless of ties to libraries or particular geographical boundaries or the ability to be preserved or "authenticated" by institutions (Skenderija 2009, p. 317 citing Matthews and Baish 2007) in the traditional library, archival, or scholarly publishing sense. Take, for example, arXiv.org, a platform for sharing pre-publication research heavily utilized by theoretical physicists—*as it existed before* it was situated within a library infrastructure at Cornell University (Ginsparg 2011) or Stack Overflow, a forum for informatics-related exchanges hosted by a commercial organization—not a library. While qualitative and quantitative methods for considering the information-related behaviors of specific communities of scientists have been developed from the perspective of many disciplines—I will examine notable IS research in the dissertation's literature review and briefly touch upon it directly below—the sheer complexity and dynamic nature of the GNAE allow much room for additional ethnographic research as well as meta-research of ethnographic studies in IS, although the latter is beyond the scope of this dissertation.

Utilization of multi-sited ethnographic techniques in this study enables acknowledgement of the network paradigm (Skenderija 1999) and allows for "thick description" (Geertz 1987) of information-related behaviors by individuals within the GNAE by:

1.) Providing a way to examine the richness, flexibility, and variety of mechanisms for scholarly inquiry in the networked environment, particularly scientific research which crosses disciplinary and institutional boundaries as well as the boundaries of formal scholarly outputs such as journal articles or books
2.) Recognizing the decoupling of information-related exploration and discovery from geographical, institutional, and format boundaries
3.) Taking into consideration broader issues that are difficult to quantify; e.g., allowing me to ask *why*–not simply *how*—scholars have begun to circumvent libraries and other institutional boundaries in the GNAE, and what this means when viewed from theoretical and applied perspectives.

Ethnographical studies combined with social network analyses have been increasingly utilized by IS researchers, as noted by Khoo et al. (2012) and Sandstorm and Sandstrom (1995), to examine information-related behaviors of individuals, groups of individuals, and communities—with notable recent mixed method analyses of communities of scientists having been performed by Pepe (2010), Pepe and Mayernek (2012), and Velden and Lagoze (2013)—however, a multi-sited ethnographic strategy has not yet been employed in IS, as of the date of publication of this thesis, to specifically examine the allotment of information-related exploration and discovery beyond official library-supported ecologies in the GNAE by scientists at multiple sites working in different disciplines. The dissertation is also, as far as I know, unique in investigating how the patterns observed using ethnographic methods might relate to broader theories about globalization, academic capitalism, (search engine) power, striving universities, and time-space compression. The dissertation also provides examples of applied ethnography, describing specific services created in response to interim findings during the research period.

I began participant recruitment and observation in for this dissertation in November 2012, several months before I began working at the Czech National Library of Technology (NTK)—a position which has allowed me to observe and interact with scientists with various disciplinary backgrounds previously unknown to me—with the majority of fieldwork having been conducted between February 2013 and September 2015. Data I gathered and interpreted over this two and a half year period include notes from in-person and remote participant observation, transcriptions of unstructured and structured interviews, and a variety of other artifacts (field notes, images, emails, web sites, and other materials). The combination of methods (research triangulation) was intended to improve data validity and generalization for the research (Clark et al. 2006). I had planned to use AQUAD 7, an open-source computer-assisted qualitative data analysis software (CAQDAS), for coding, linking, mapping, and annotating of qualitative data (Huber and Gürtler 2014), but in the end, found the creation of my own system for doing this work to be more useful and efficient—notably in developing with small datasets which I could analyze, manipulate, graph, and more easily share with others, should the need arise.

1.3 Structure of the Dissertation

Following this introduction, I describe in Chapter Two the process in which I approached the central research questions—the theoretical context in which the dissertation is placed. I review the literature which is relevant for conducting a multi-sited ethnographic study in IS. I examine the concept of the GNAE itself to begin my analysis—the GNAE being the global infrastructure within and upon which the scientists I observed conduct their work. I then discuss why qualitative methods and the employment of multi-sited ethnographic strategies are relevant to the central research questions and how ethnographic research has previously been utilized in IS research. Chapter Three discusses how I designed and implemented my project in relation to the central research questions and theoretical literature. In Chapter Four I *richly describe* the data gathered throughout the research process. In Chapter Five, I provide a discussion of the findings and relevance of these data, including practical application of findings to service design. I concisely conclude and summarize findings of the dissertation in Chapter Six.

Chapter 2 Literature Review and Theoretical Framework

2.1 Overview

In this chapter I conduct a literature review. I first define the GNAE as it is relevant to this study and then discuss my research methodologies according to theory and the work of previous scholars, including prior research in IS and anthropology, where appropriate.

2.2 Global Networked Academic Environment (GNAE)

As mentioned in the problem statement, I utilize the term GNAE to encompass previous descriptions of the networked environment referred to since the advent of the Internet with various terms ranging from *fluid hypersphere* (Tsekeris 2008) to *dataspace* to *e-Infrastructures (e-IS)* and beyond (see notably Borgman et al. 2012; Pollock and Williams 2010; Skenderija 2008; Willcocks and Whitley 2009). The GNAE specifically refers to the complex bricolage of network infrastructures, online information resources and tools scholars use to perform exploration and discovery today (i.e., global academic e-IS or, according to Edwards et al. (2013), knowledge infrastructures). In this way, the GNAE is the ever-dynamic global "stage" *upon*, *with, and through which* the individuals I observed throughout the research process interacted when conducting scientific exploration and discovery. As such, it is important to describe its basic characteristics, as defined by prior researchers, in order to understand my selection of research methodologies as well as my later research findings.

2.2.1 GNAE Characteristic One: Concurrently Virtual/Global and Physical/Local

The characteristics of the GNAE played a crucial role for me in selecting a multi-sited ethnographic approach incorporating visual ethnographic techniques as research strategies, because they are particularly well-suited, as will be seen below in Section 2.3.4.2, to examining the "performative reality-structure" (lifeworld) of the local and global in day-to-day scientific work, in which the "virtual world is no less real (or less promising) than 'real life' (off-line world)" (Tsekeris 2008, n.p.). In other words, I made the crucial assumption in this dissertation that, in the GNAE, the virtual (global) world is as real as the physical (local) world and, as such, designed my research to include participants from various geographic sites as well as data gathered both in-person and online (see Chapter Three, Research Design).

2.2.2 GNAE Characteristic Two: The Definition of Information Extends to Include Virtual Interactions

I additionally took the existence of the GNAE, with emphasis on *networked environment*, as a theoretical given, building upon Czech philosopher Miroslav Petříček's media theory and post-structuralist theory as examined by scholars ranging from Deleuze to Baudrillard, summarized in English by Skenderija (2008, p. 6), where the global network:

> Constitutes a virtual order which is not only a digital world of radical speed and hyperconnectivity—but also a place where each user at each interface within the network is at the same time both an interpreter AND an interconnected creator of the network.

In the GNAE, the definition of information is therefore extended beyond *information-as-thing* to include, as Buckland (2012) cites Furner (2004)*, information-as-knowledge* and *information-as-process* not only in the physical realm but also in a dynamic virtual realm—as Skenderija (2008, p. 6) puts it:

> Information here is no longer simply an object conceived from the point of view of an isolated subject sitting at a computer screen, but instead, every interaction itself becomes a virtual configuration of a certain type knowledge, and this virtual configuration itself, now understood as information, becomes a constitutive element of the network itself.

This served as an important consideration for my project in terms of research design, data collection, and interpretation—virtual interactions must be included to accurately capture all kinds of possible information in the network.

In other words, the GNAE here can be considered to be a conceptual plane of consistency in the sense of Deleuze and Guattari (1987, p. 70), upon which the aurae of various underlying *information strata (i.e., content and expressions of information of any kind)* race or dance in a "continuum of intensities."

2.2.3 GNAE Characteristic Three: It May Include Multiple Networks

Edwards et al. (2013, p. 5) build upon prior science and technology studies and define knowledge infrastructures as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds." In this dissertation, I purposefully employed the singular term GNAE instead of knowledge infrastructures from the outset of my research as a shorthand in order to broadly encompass any possible various knowledge infrastructures and network interactions I might encounter during the course of the study, realizing in advance that 1) I could not predict what knowledge networks or interactions I might observe and 2) that the GNAE includes rapidly changing:

> [E]cologies or complex adaptive systems [consisting] of numerous systems, each with unique origins and goals, which are made to interoperate by means of standards, socket layers, social practices, norms, and individual behaviors that smooth out the connections among them. (Edwards et al. 2013, p. 5)

This was a key assumption as I designed an approach to exploratory RQ1 (*If and how non-institutional/informal information-related exploration and discovery is occurring beyond official library-supported mechanisms in the GNAE)*, because it meant I would need to develop a method for understanding these *dynamic networked information ecologies* in which scientists conduct their research. With the a priori assumption that the GNAE exists, I could therefore assume I would be able to observe its constituent ecological components—ever-changing as they may be—at points of virtual and physical intersection of actors (individuals, artifacts, institutions) with and within networks.

2.3 Research Methodologies and Theory

2.3.1 Qualitative versus Quantitative Research Methods and IS

In a survey of mixed method research (MMR) in library and information science (LIS) journals, Fidel (2008)—acknowledging the lack of consensus by social scientists on a definition for qualitative methods—creates a concise operational definition of qualitative versus quantitative research methods: the former creates text (including multimedia) as an outcome; the latter, numbers. The methodological terrain is obviously much more complex than this, as Fidel recognizes—with polarization between advocates of the two methods ebbing and flowing according to discipline or historical period. Sandstrom and Sandstrom (1995, p. 180) remind IS researchers that "[n]othing in ethnographic research militates against quantification of data"—including descriptive quantification. I would add that in a multi-sited ethnography, *deep description* of data sources themselves should also be considered for some projects (see Section 4.5.3).

I will not provide here a more detailed description of qualitative or quantitative research methods; as Cibangu (2013) notes, textbooks describing them abound. A recent textbook including MMR is Creswell (2014).

When it comes to IS, as noted in the introduction to this dissertation, the wide variety of possible research questions enables researchers to select the most appropriate method in relation to the research question(s) at hand. According to Fidel (2008) and as illustrated by Creswell (2014), the choice is no longer simply binary (quantitative or qualitative); MMR is already the third pillar of research approach in many social sciences (e.g., in the field of human development, as reviewed by Yoshikawa et al. 2008), and might be particularly helpful in interpreting and explaining quantitative research results—*adding (con-)text to the numbers*, one might say using Fidel's operational definition.

Cibangu (2013) criticizes the quality of qualitative work in IS, noting the dearth of theory creation, its lack of ties to other bodies of knowledge, and misunderstandings of essential concepts. He identifies a need to improve accessibility to research in terms of more fastidious abstracting and creation of titles based on his survey of qualitative IS articles dating from 2001 to 2011.

2.3.2 MMR and Network Science in IS

Because RQ1 involves observation of information behaviors within the GNAE, I considered utilizing MMR methods, having been made aware in 2012 by a contact, an established theoretical physicist at an American research university, of recent mixed research conducted by Velden and Lagoze (2013) on networked communities of scientists. Velden and Lagoze's research, in turn, led me to notable MMR work about networked scientific communities by Pepe (2010) and collaborative value production (Pepe 2011). These works embraced qualitative methods to *add context* and to inform what are primarily quantitative network analyses.

Network science (Barabasi 2002; Börner et al. 2007) is often encountered in IS in the metrics branch of inquiry (bibliometrics, scientometrics, infometrics, webometrics, e-Metrics) and often utilizes data sources such as citation information: rich, varied, and relatively easy to harvest. These data sources can be graphed because they have properties such as nodes, edges, and distributions; the resulting graphs (visualizations) can then be interpreted.

During the research process, I was encouraged several times by my key research participants to consider an MMR approach—and at one point in mid-2013, I nearly veered (in the "going native" sense) into including a quantitative component to my research design. However, my advisor reminded me about the difficulty of obtaining a representative sample in relation to RQ1: "I am concerned that your data sources are too thin" (M. Seadle 2014, pers. comm. 23 August), and he was right—because during the course of my project, I encountered a sampling frame problem in the quantitative sense, which was the source of my "thin data": I was dealing with a hard-to-reach population. But as will be seen in Chapters Four and Five, this does not preclude the illustrative use of network and other visualizations to assist in understanding theoretical concepts and patterns of information use.

2.3.3 SS/RDS and Hard-To-Reach Populations

I will discuss and describe my research population in more detail in Chapters Three and Four, but will comment here on the theoretical aspects of the sampling frame in relation to RQ1: *According to the multi-sited ethnographic analysis of scientists participating in this study—individuals conducting research in various disciplines at different institutions in several geographical locations—is there evidence indicating a significant allotment of non-institutional/informal information-related exploration and discovery occurring beyond official library-supported mechanisms in the GNAE?.*

Answering this question in quantitative terms would require a different approach to the sampling frame: RQ1 would have to be modified in order to obtain a random, potentially representative sample population *of something* in order to explore significance in the statistical sense. In doing so (for example, by limiting the study to a particular scientific discipline or institution as is the case in Pepe 2010), one could employ snowball sampling/respondent driven sampling (SS/RDS) in the sense of Goodman (1961), where an initial random sample from a given finite population names additional population sample members, who in turn name more sample members in "waves" of naming. This would, however, require that the initial sample be truly random in the statistical sense.

For this dissertation, however, I wanted to be more exploratory in my approach, and did not wish to limit the population strictly—nor did I wish to limit my definition of GNAE to a particular available data set(s). I wanted to be surprised about my observations of GNAE constituency instead of limiting my decisions about it in advance.

I also discovered several months into the research process that for my particular time-limited research project and individual circumstances—conducting my research part-time without full-time access to a particular organization and/or research group—unlike Allard et al. (2009), whose research team included fifteen project team members in addition to three authors and the support of The Institute of Electrical and Electronics Engineers

[IEEE] and six companies in selecting research participants—actively-working research scientists comprise a *hard-to-reach population* as defined by Goodman (2011, p. 350): even if a statistically-sound sampling frame might be able to be developed, it would be "too difficult and/or too expensive to use it in order to obtain a random sample of the population." To utilize quantitative methods in this study, even if I were to limit my sampling frame to the relatively small international community of cheminformatics practioners in the GNAE, I would have to develop and define a list of all possible practitioners *around the world*—i.e., I would require a multistage/clustering technique (Creswell 2014, p. 158)—in order to even begin thinking about a truly random GNAE sample with which I might begin to analyze. While this is not an impossible task (although any such list would be outdated the moment it is completed) and while it is a task that might even be able to be partially automated, as a lone researcher working within a dissertation timeframe of three years, I concluded it was *too time expensive* and made a methodological tradeoff: I did not want to sacrifice the ethnographic aspects of this dissertation in exchange for walking a purely quantitative or MMR path—the latter which, however, may hold promise for future IS research, as illustrated by Velden and Lagoze (2013) and Pepe (2010). For me, investigating the "search for meaning" in relation to RQ1 became more important than "experimental science in search of law" (Geertz 1987, n.p.), as I address in the next section and later in terms of research design.

In ethnographic studies, a hard-to-reach participant group need not be completely random and the definition of snowball sampling is not taken literally; a recent example of this is found in Walsh (2014), who—despite providing financial incentives to potential participants in her research into the Facebook behaviors of adolescents—employed a broadly-defined "snowball" technique when other attempts to recruit participants failed: she resorted to personal and professional connections to gain access to a hard-to-reach community. As will be seen in Section 3.3, this was also my case, with one exception.

2.3.4 Ethnographic Strategies

Ethnographic studies share, regardless of particular ethnographic strategies employed in relation to different research questions, an intensive interest in observing human behavior. Khoo et al. (2012) and Riemer (2008) provide concise histories and overviews of ethnographic methods. Ortner (2007, p. 788) notes in her obituary for Clifford Geertz:

> Our job as anthropologists is to get at the meanings that shape and inform all of social life. It is also about how we may and must go about uncovering such meaning, namely through reading social life as if it were a text, or as a text, to be interpreted, on analogy with the interpretation of literary texts" [the so-called interpretive approach (Martin, 1993)].

According to Sandstorm and Sandstorm (1995, p. 164), "potential uses of ethnographic methods are infinite, limited only by the scholars employing them."

The number of ethnographic studies in IS has been growing in recent years (see Section 2.3.4.3 below), including applied studies addressing information system design issues—Crabtree et al. (2000) and Crabtree (2012), for example, proposed ethnomethodologically informed ethnography as a way to inform systems design—as well as perspectives of library users à la Gabridge et al. (2008). Applied ethnographic studies also appear in in other social science disciplines such as business (e.g., Boden et al. 2001; Mills and Ratcliffe 2012; Prior and Miller 2012) and software development/computer-mediated communication/computer-supported cooperative work (CSCW) (Barry 1995; Beynon-Davies 1997; Beynon-Davies et al. 2000; Blomberg and Karasti 2013; Jirotka et al. 2013; Garcia et al. 2009; Kjeldskov and Stage 2012).

I will discuss the theoretical literature relevant to specific aspects of research design and data interpretation directly in relation to each individual component of the study in Chapters Three and Five.

2.3.4.1 Prior Ethnographic Research in IS

High quality reviews of ethnographic work in IS have been produced in the past twenty years, notably by Sandstorm and Sandstrom (1995), Khoo et al. (2012), and Boukacem-Zeghmouri and Schöpfel (2013).

Sandstrom and Sandstrom (1995, p. 191) provides potential ethnographers with five qualitative research design assumptions which are often neglected by IS researchers. Although twenty-years old, the article remains highly relevant, stating:

> The goal in [library and information] LIS inquiry is to increase knowledge about all aspects of human information behavior, not only for its own sake but in order to provide a basis for solving real-world problems.

The authors also caution IS researchers to be aware of trends outside of IS, *if* IS researchers strive to attain broader relevance and/or resonance in the social sciences.

Khoo et al. (2012) provides a comprehensive review of ethnographic research studies in (L)IS between 1980 and 2011 and includes historical and methodological synopses as well as a comprehensive bibliography of articles using a discrete set of top English language LIS journals; journals in other languages are absent from the review. Two works mentioned by Khoo et al. which deal with information behavior in academia include: Barry (1995), which touches upon the contextual issues related to introducing IT to the academic environment, but without a focus on scientists; and Future (2001), which explores use of eJournals by biomedical scholars. However, as Velden and Lagoze (2013, p. 1) note—citing Gläser (2006)—"comparative studies of scientific fields with an ethnographic depth of understanding of research practices and social behaviors are rare."

Indeed, when writing this dissertation, I was unable to locate prior IS studies which specifically stated they were using multi-sited *ethnographic* strategies in order to specifically investigate exploration and discovery by scientists *in different disciplines at multiple sites* (RQ1); works by Allard et al. (2009), Tenopir and King (2004), and Levine et al. (2011) focus on engineering only. And while Tenopir and King (2004, Chapter 11) include comparisons with other scientific areas, they do not refer to their methods specifically as being multi-sited ethnographic.

Boukacem-Zeghmouri and Schöpfel (2013) observed science, technology, engineering, and mathematics (STEM) researchers in five disciplines at five universities in France over a four-year period, but they do not explicitly state they utilized multi-sited ethnography as a research strategy. They provide a review of information-behavior related studies since 2000—some of which include ethnographic components—noting (pp. 142-143):

> The results provide more or less anecdotal evidence, for example a patchwork-like description rather than consistent data on information-seeking behaviour in different scientific communities. In other words, it is not possible, at least for the moment, to draw a consistent picture of specific heuristic patterns related to digital information.

Their key findings include observations of local/global duality by researchers, standardization of search engine use (Google), and a "total absence of the concept of a library [in the physical and traditional sense] from the researchers' discourse" (p. 147).

Another study of researchers at different sites in the same country is Ellis et al. (1993), who interviewed research physicists and chemists at Manchester University and the University of Sheffield. They analyze the results of their interviews according to "the constant comparative method of analysis as outlined by Glaser and Strauss" (p. 357) and describe how the scientists (and social scientists from a previous study) conducted information seeking according to eight categories: starting, chaining, browsing, differentiating, monitoring, extracting, verifying, and ending (p. 359).

As noted in Robbins (2011)—although this article comparing the information behaviors of engineering faculty at twenty institutions is not ethnographic—many other ethnographic studies in IS are institutional- or resource-specific (see also Boukacem-Zeghmouri and Schöpfel 2013). I turned to the anthropological and sociological literature, as advised by Sandstorm and Sandstrom (1995), for a better theoretical understanding of why so many ethnographies in IS focus on individual settings.

2.3.4.2 Multi-Sited Ethnography, including Virtual Components

Any kind of global/local discourse has only recently and tentatively entered qualitative studies in IS, and most often it appears in a kind of essay format (Anderson 2014; Anderson 2015; Frohmann 2013; Ye et al 2013). However, in quantitative metrics studies in IS and in other social sciences, geographical comparison (in metrics studies) and globalization (in other social sciences) have been topics of analysis for some time. Regarding multi-sited ethnographies of sciences or scientists, Escobar et al. (1994) recalls, for example, Margaret Mead's cybernetics research program in the mid-twentieth century as well as Arjun Appadurai's "global ethnoscapes" within the context of a global cyberculture, and notes such projects involving a multi-sited perspective may not be different than traditional enthographies in that they still focus on cultural diagnoses and "emerging practices and transformations associated with rising technoscientific developments" (p. 216).

I would argue that qualitative IS research has only begun to come to terms with the contextualization of its various research activities; its focus on individual products, institutions, behaviors, and collaborations must be expanded in order to grapple with the complexity of the GNAE and to provide appropriate responses to the applied aspects of the discipline (Luft, 2015)—and if the phenomena being observed are in "cyberia" (i.e., the superset of the GNAE, the global network), research strategies must take into account virtual, multi-sited global ethnoscapes. Multi-sited ethnographic studies which include virtual components are one way to do this. And as Skågeby (2012, p. 325) notes, "[i]n many ways, online ethnography is no different from traditional ethnography—it is however adapted to the circumstances of online communication and communities."

Marcus (1995, p. 95) provides an essential explanation of the emergence of multi-sited ethnography, a shift away from single-site ethnographies to:

> [M]ultiple sites of observation and participation which cross-cut dichotomies such as the 'local' and the 'global,' the 'lifeworld' and the 'system.' Resulting ethnographies are therefore both in and out of the world system.

Multi-sited ethnography is, in my opinion, perfectly suited to studies of information behaviors within the GNAE; it offers a way of examining "the circulation of cultural meanings, objects, and identities in diffuse time-space" (Marcus 1995, p. 96). Marcus (p. 99) notes that by mapping terrain in a multi-sited ethnography:

> Its goal is not holistic representation, an ethnographic portrayal of the world system as a totality. Rather, it claims that any ethnography of a cultural formation in the world system is also an ethnography of the system, and therefore cannot be understood only in terms of the conventional single-site mise-en-scene of ethnographic research, assuming indeed it is the cultural formation, produced in several different locales, rather than the conditions of a set of subjects that is the object of study.

In terms of RQ1 (*According to the multi-sited ethnographic analysis of scientists participating in this study—individuals conducting research in various disciplines at different institutions in several geographical locations—is there evidence indicating a significant allotment of non-institutional/informal information-related exploration and discovery occurring beyond official library-supported mechanisms in the GNAE?),* I therefore observed scientists at different locales interacting with a *cultural formation*, the aforementioned GNAE, and its constituent information ecologies—and then considered and traced patterns of information circulation, of *information resource use (i.e., exploration and discovery pathways)* at a comparative, multi-sited level. Marcus (1995, p. 112) calls this "sorting out the relationship of the local to the global."

Marcus' impact to date on IS research has been non-existent for all intents and purposes, although his work is cited broadly in other social science disciplines. In a review of Web of Science citations for Marcus' seminal article conducted on 13 July 2015—the 1995 article was not indexed by Scopus at the time of writing this dissertation—I was unable to identify any indexed trace of Marcus' work in the journals reviewed by Khoo et al. (2012). I retrieved eight articles categorized under the topic INFORMATION SCIENCE LIBRARY SCIENCE, none of which were from journals included in Khoo's review (Appendix A). Six hundred seventy other citations were in journals spanning the social sciences spectrum.

More generally, "post-Khoo" ethnographic research (MMR research has been listed in Section 2.3.2; see also Jamali and Asadi 2010) in IS has tended to introduce ethnographic concepts or to *follow the technology* and, for the most part, might be considered applied ethnography. Examples of post-2012 ethnographic work in IS include: McEwen (2012) (mobile phones and hybridized information centers); Emary (2014), Rosenblum (2015), Seeliger (2013), Haas (2014) (ethnographic methodologies and concepts); Rhinesmith (2014) (cloud computing implementation for a community-based organization in Illinois); Hartel and Thomson (2011) (visual research in anthropology and sociology); Lingel (2015) (urban information behaviors for newcomers to New York City). Boukacem-Zeghmouri and Schöpfel (2013) and Rowlands et al. (2008) observe information behavior in relation to Google. One notable dissertation utilizing ethnographic strategies is Zhou (2010), who describes information flows within a hospital ward. While not theoretical in emphasis, her work does "thickly describe" the information flows in her research setting.

2.3.4.3 Visual and Online Ethnography

My project necessarily included visual and online components (see Chapter Three: Research Design). All informants I observed, even if experimentalists, conducted exploration and discovery at computer workstations or laptops; I did not observe any participants using mobile devices for this purpose. In order to see what is really happening as scientists interact with the GNAE, one must find a mechanism for entering the virtual world and observing information behaviors *at the point(s) of interaction with the network*. These behaviors are difficult to observe conventionally because of the speed of work with information resources online. All participants I observed for this project worked primarily in silence, the silence being interrupted only by keyboard and mouse clicking and an occasional casual discussion regarding where to go to lunch or have a beer. In such settings, asking participants about their work online would interrupt their natural research process and risk the Hawthorne effect (Allard et al. 2009, p. 453) and discovery and exploration of information as represented on a physical interface simply takes place too quickly to be recorded manually. Therefore, I initially envisioned participants conducting recording of their video screens while conducting information-related exploration and discovery and turned to the visual ethnography literature to learn more about the scholarly context for such techniques.

Figure 1: Workplace of Condensed Matter Physicist, Czech Republic.

Regarding visual ethnography and IS, Hartel and Thomson (2011), in their methods-oriented paper, review the relevant historical literature and discuss the validity of using visual material as part of an ethnographic study. They propose the use of images in IS, cite four examples, and describe how visual ethnography might especially inform research into "immediate information spaces" (p. 8).

Observations of virtual spaces, so-called interface anthropology, have been employed to investigate human-computer interaction for the past two decades (Escobar et al. 1994); this is important in the context of this dissertation because it allowed me to include analyses of visual materials as an expanded "awareness context" (Hine 2006). Skågeby (2012) notes how ethnographers have always used document collection in physical environments to supplement observations and interviews; in the online context, document collections (textual documents and multimedia materials such as videos and photos) are "perhaps [the] most important form of data gathering" (p. 325). Whittle (2000) discusses why ethnography has necessarily moved online and discusses the contours of network ethnography, including its limitations. He proposes richly describing qualitative data in network form, borrowing concepts from quantitative network science: “A thick description of a network has to illustrate and illuminate the nodes, links, and flows, the structuration of the network” (‘From the Field to the Net,’ para. 11). I do utilize network visualization by illustrating use of information by participants in this study in Chapter Five.

2.3.4.4 Ethnographies of Science, Scientists, or Technology

I will not provide an exhaustive list of all possible ethnographies of science, scientists, or technologies here, but will highlight: 1) notable IS ethnographic works including science and technology—not using Khoo’s journal set but rather Web of Science’s broader INFORMATION SCIENCE LIBRARY SCIENCE topic identifier, and 2) works from other disciplines which I found relevant in writing this dissertation.

In IS, in addition to the MMR network analyses listed in Section 2.3.2, Allard et al. (2009) note that studies of the information behaviors of engineers date back half a century, with a summary of findings available in Tenopir and King (2004). The aforementioned Boukacem-Zeghmouri and Schöpfel (2013) utilize ethnographic observations and stated this explicitly in their book chapter abstract. Ellis et al. (1993) compare the behaviors of physical and social scientists and stated their work is qualitative and “based on the grounded theory approach” (p. 356). Jamali and Asadi (2010) also supplement survey data with semi-structured interviews of physicists and astronomers.

More broadly in the social sciences, Hine (2006) remarks ethnographic methods—particularly observations of laboratories—have become common in the sociology of scientific knowledge (SSK), with Latour et al. (1986) being a seminal work.

Recent social science (e.g., sociology or anthropology) dissertations which explore questions relating to science or technology and utilizing ethnographic strategies without MMR include Luo (2014) (first-year engineering students and social media technologies), Walsh (2014) (Facebook and adolescents), and Dick (2010) (sociology of superstring theory). Recent dissertations from (L)IS schools include the aforementioned Rhinesmith (2014) and Zhou (2010).

2.3.4.5 Limitations and Critiques of Multi-Sited Ethnographic Strategies

I will discuss limitations of the individual components of my research design in relation to theory in Chapter Three.

Regarding multi-sited ethnographic strategies more generally, Marcus (1995, p. 99) describes possible limitations as "methodological anxieties." He notes that multi-sited strategies might test the limits of ethnography itself by "stretching" older definitions of single-sited ethnography (p. 99). He states that the "mystique and reality of conventional fieldwork is lost in the move toward multi-sited ethnography," but counters that "multi-sited ethnographies inevitable are the product of knowledge bases of varying intensities and qualities" (p. 100) and reminds researchers *to know the language* of those they observe in order to avoid quality issues. Finally, there might be a "loss of the subaltern" perspective (p. 101); this can be addressed by adding a subaltern focus to some area of a project—but even if not, including a comparative dimension in a research project can counterbalance this:

> Comparison emerges from putting questions to an emergent object of study whose contours, sites, and relationships are not known beforehand, but are themselves a contribution of making an account that has different, complexly connected real-world sites of investigation. The object of study is ultimately mobile and multiply situated, so any ethnography of such an object will have a comparative dimension that is integral to it, in the form of juxtapositions of phenomena that conventionally have appeared to be [or conceptually have been kept] 'worlds apart.' Comparison reenters the very act of ethnographic specification by a research design of juxtapositions in which the global is collapsed into and made an integral part of parallel, related local situations rather than something monolithic or external to them (p. 102).

Since 2005, defenders and detractors of Marcus' multi-sited strategies have emerged. Candea (2007, p. 180) notes there may be a potential loss of arbitrariness, defined as "space which cuts through meaning," if fieldwork locations are completely unbounded. Nadai and Maeder (2005, p. 3) also address the importance of defining a field and stated it must be attuned to the particular research question; however, they feel the "fuzziness" of fields—fields without clear boundaries—can be counteracted by a "theoretical clarification of the object of study…such a theoretical framework can then serve as a compass for research."

Falzon (2012) summarizes threads of criticisms between 2005-2012, including lack of depth (p. 7), abdication of ethnographic responsibility (p. 10), and "Latter-day Holism" (p. 12). In response to these, he describes a second generation of multi-sited ethnography characterized by spatial depth (to counteract shallow description) and guiding research by thorough awareness of prior scholarly literature (to ensure/reveal ethnographic partiality). To counteract holism charges, Falzon (2012, p. 13) advocates making sure that multi-sitedness “means not just sites, but spatialized [cultural] difference—it is not important how many and how distant sites are, what matters is that they are different.”

Hine (2007) praises the innovative capacities of multi-sited ethnography and encourages a mixture of on- and offline observations as a way to counteract a potential lack of depth in virtual multi-sited research. She reminds researchers to remember David Hess’ definition of good ethnographies, which include “a deep knowledge of the field of endeavor they cover [and which] contain surprises and subverts the obvious” (p. 661). She encourages an “adventurous spirit” in research design—to not cast aside methodological traditions in the past, but rather to adjust them to better-fit the realities one encounters in fieldwork (p. 667)—even if this means abandoning the methodological label “ethnographic” in a particular research report.

In this dissertation, I specifically designed my project to counteract possible limitations of the multi-sited method, taking an adventurous approach in the sense of Hine (2007) and addressing the aforementioned methodological concerns to the best of my ability, as I describe in Chapter Three.

2.4 Summary of Literature Review

This dissertation embraces multi-sited ethnographic strategies as a way to appropriately study the exploration and discovery pathways of scientists in the GNAE. I assume that the GNAE does exist according to post-structuralist theory (an in-depth philosophical inquiry into this topic, however, is beyond the scope of this dissertation), and also assume that visual and virtual ethnographic methods can be applied to my qualitative research questions in IS. Though grounded in theory and designed accordingly, this dissertation is purposefully exploratory and adventurous in the spirit of Marcus (1995) and Hine (2007) and additionally embraces “thick description” in the networked environment by exposing otherwise invisible elements of information-related behaviors in the GNAE in relation to my primary research questions.

Chapter 3 Procedures and Research Methods

3.1 Research Design

3.1.1 Summary

I designed this project from the very outset to be repeatable over different populations and also theoretically applicable to future questions about information behavior, having been frustrated by the patchwork flavor of prior ethnographic studies in IS regarding scientific information behaviors (Boukacem-Zeghmouri and Schöpfel 2013) and a dearth of useful prior multi-sited ethnographic research models in IS. I did not, for example, locate one single IS ethnographic analysis that offered prior examples of codes describing the modes or *models of production* of information identified (i.e., how a resource is produced; for the origins of this term, see Russ 2012, pp. 137-163).

Allard et al. (2009, p. 448) and Hemminger et al. (2007) provide a helpful description of their data interpretation and coding processes and served early on as a model for me in designing my research, but they focus on *communication events and formats* rather than on *who produces or facilitates access to* information resources. One encounters this also in earlier studies such as Brown (1999)—from a historical perspective, her analysis is interesting and useful, but her project limits itself to a narrow single-sited survey approach. To me, many qualitative research designs I contemplated when preparing this study seemed like methodological straightjackets, lacking global perspectives and limiting themselves to pre-network paradigms, as seen in the literature review above.

Because of this, I designed a set of multi-sited exploratory ethnographic strategies to address my research questions. The research methods and procedures I describe below could easily be repeated over many different kinds of populations, and resulting data about information behavior patterns could be systematically compared using the same or similar coding techniques. Many interesting research questions could be considered using these and similar methods and even by complementing the qualitative dimensions with related quantitative or MMR work. The field is, in some ways, wide open. Future research is limited only to our "imaginaries" as IS researchers. I feel strongly we need to broaden our research horizons, to become more stringent about designing our research, and to improve theoretical and methodological linkages to other disciplines. In other words, to overcome the current *potpourri-like* nature of qualitative IS research and its tendency to follow buzzwords and trends which have emerged in other disciplines, we need to become more systematic and thorough—although this does not mean we cannot still be eclectic, as I will discuss in the conclusion to this dissertation.

Because my research design is exploratory, I will describe the design process in detail.

3.1.2 Early Stages: Creating a Research Compass

I began the design process in summer 2012 by sketching my thoughts about possible research questions on paper after conducting an initial literature search about research methods in IS. I wanted to learn more about the information-related behaviors of scientists, anticipating my enrollment as a doctoral student and preparing for a role in a science and technology library beginning in March 2013. From my previous professional experiences, I felt I had a solid understanding of the library-related dimensions of scientific information use, but I was unaware of what was happening *beyond the library paywall*. I therefore purposefully wanted to take off my "information professional" blinders and observe the GNAE through the eyes of scientists themselves—exactly which scientists, I did not know yet. The following sketch represents the "foreshadowed problem" of this study (Riemer 2008) and an initial conception of my research interest.

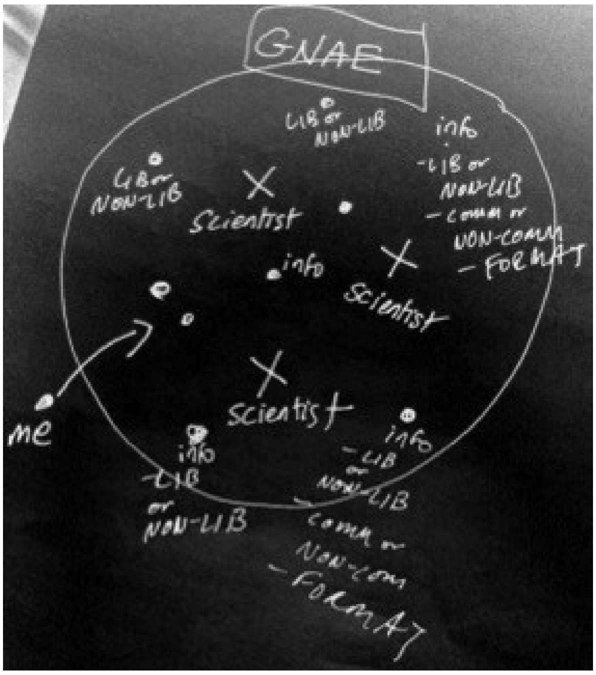


Figure 2: Initial Research Design Sketch.

This simple sketch proved remarkably valuable throughout the evolution of the study by providing me with a sense of *field boundedness* throughout the project. It served as a kind of compass to guide me as I adjusted my research approach to fieldwork experiences and findings across time and space. Instead of studying the GNAE itself (an enormous fieldsite) or defining within it a random sample population of scientists I might study in the quantitative sense prior to conducting my fieldwork, I posited I would be able to—with a rich variety of in-person and virtual ethnographic methods—be able gather data from *even a small and not necessarily randomly-selected population of (yet-unknown) scientists (n>=1)*. I additionally postulated that whatever data I gathered during my fieldwork with this broadly-defined population—defined initially simply *as one or more actively-working scientists,* whose information-seeking behaviors (cultural patterns) were, beyond the context of JSTOR, previously unknown to me—would provide me with valid and reliable insight into my exploratory primary research questions. In other words, even if my population ended up being n=1, I would be able to say something about the non-library-supported information-related behaviors *of that observed scientist (RQ1), observe patterns of behavior (RQ2),* and compare findings to theory. If the number of individuals I observed ended up being n>1, I would be able to additionally compare the behaviors of these individuals *with one another* in addition to theory. In this way, the more individuals observed therefore results only in more depth and complexity in terms of comparative understanding in relation to exploratory RQ1 and RQ2. And multi-sitedness, then, just *added another layer of possible comparison* (i.e., depth and complexity) into the study, without making any claims about the universal applicability of research findings, thereby avoiding holistic concerns regarding multi-sited ethnographical strategies. Boundedness, in this manner, is not defined by specific fieldwork sites or populations observed (as in some traditional single-sited ethnographies)—it is the *primary research questions which remain constant* and provide boundedness. The research results of such a project should, therefore, be reliable and valid as long as they are not generalized beyond the research questions at hand.

With my central research questions in place, I then generated an initial list of core conceptual coding categories which might realistically be applied to any data gathered in the field; the following table defines the categories seen in my initial research design sketch (Figure 2) in relation to my central exploratory research questions.

Table 1: Initial Core Conceptual Coding Categories, Definitions, and Relations to Primary Research Questions.

Category	*Definition*	*Related to RQ?*
LIB	Information in the GNAE accessed by a research participant behind the library paywall for an institution to which a participant belongs (i.e., subscription database or other resource provided by a university library)	RQ1 (library versus non-library use—central information phenomenon I wished to observe)
NON-LIB	Information in the GNAE accessed by a research participant beyond the library paywall for an institution to which a participant belongs (anything)	RQ1 (library versus non-library use—a central phenomenon I wished to observe)
COMM	The entity who/which hosts the presence of the information in the GNAE is a commercial entity (person, institution, provider of other kind)	RQ2 (patterns; perhaps useful for interpretation of data down the road relating to theories about globalism and academic capitalism which I had read and found interesting—response to the critique of possible loss of the subaltern [Marcus 1995])
NON-COMM	The entity who/which hosts the presence of the information in the GNAE a non-profit entity (person, institution, provider of other kind)	RQ2 (same as above)
FORMAT (with attributes)	Could possibly include attributes to identify formal versus informal data type (e.g., print book "authenticated" by a trusted institution versus lab notes published on an individual's blog)	RQ1 (informal versus formal information component of this research question)

Creating this initial list of codes even before entering the field gave me reassurance that my research questions could be answered. I expanded the complexity of codes and relationships between them over the research period and will discuss this in Section 4.3.

At this early point in designing the project, I was most influenced by charts of communications events observed by Allard et al. (2009) and Hemminger et al. (2007), as noted above. As such, I considered what data-gathering techniques might feasibly be employed in the field even before reaching out to potential research participants. From the very beginning, I included both physical and virtual data sources and anticipated multiple sites for research, knowing I would be moving from the United States to the Czech Republic. I did this instinctively and admittedly did not discover Marcus (1995) until the last year of my study, at which point I was then able to anchor my instinct in theory.

Table 2: Initial Data Source Ideas.

Data Gathering Technique or Source	*Data Type*	*Anticipated Data Use*
Video Screen Captures	Video	To observe virtual information interactions of scientists with the GNAE
Skype Interviews	Video or Skype text transcript	To discuss information-related behaviors and clarify behaviors with research behaviors remotely
In-Person Interviews	Text (transcription of text or audio notes)	To observe informants in their own work settings, in order to gain visual data (e.g., images of workplace settings) and for possible future comparison with Skype interviews
Wikipedia Talk Logs	Text	To observe how scientists discuss and develop important concepts online
In-Person Observation	Text (field notes)	To supplement and enrich the virtual observations

In the end, sources of data gathered in the field varied from this initial set (see Chapter Four) and I was unable to convince any research participants to provide me with video screen captures of their research activities, although I identified an easy-to-use freeware software for this purpose. As an alternative to video, static screen captures replaced videos—and surprisingly, the screenshots ended up being simple but extraordinarily rich sources of data.

3.1.3 Defining the Field: Virtual Interactions as Key Observation Points

Falzon (2012, p. 9) refers to Bronisław Malinowski's requirement for participant observation in ethnography as the "main portal to the native's point of view" and states that this "perhaps constitutes the strongest case for multi-sited ethnography…If our object is mobile and/or spatially dispersed, being likewise surely becomes a form of participant observation." For my study, this of course meant I must include multi-sited conventional observation components (to add horizontal and vertical depth to the study) once I found research participants (more about this in Entering the Field, Section 3.3). But what does *participant observation* mean in the virtual, networked sense in relation to my central research questions?

Recall Skenderija (2008) from the theoretical discussion of the GNAE (Section 2.2.2.2 above): within a network, virtual interactions of people with information also constitute information and therefore can be considered to be, at the moment they exist, observable components of a network. Because of this, it becomes possible to consider *virtual interactions as key points of observation* of how scientists (research participants) are working with information within the GNAE.

In designing this study, this meant I needed a mechanism for collecting data about virtual interactions. As noted above, I had initially thought this would be video. In the end, because of reluctance on the part of participants to use video, I asked my key informants to create screenshots of their information-related behaviors defined by the participants themselves. This provided me with their emic perspectives on information use. I supplemented this data with in-person observation, interviews, follow-up meetings and emails, and a pre-interview questionnaire. The screenshots provided me with snapshots of virtual interactions, and these virtual interactions can be mapped and studied in qualitative network form (Whittle 2000), with scientists and information resources viewed as "nodes," each with their own attributes. The resulting patterns can then be analyzed and interpreted (even if n=1 if virtual interactions observed>=1) in order to answer RQ1 and RQ2.

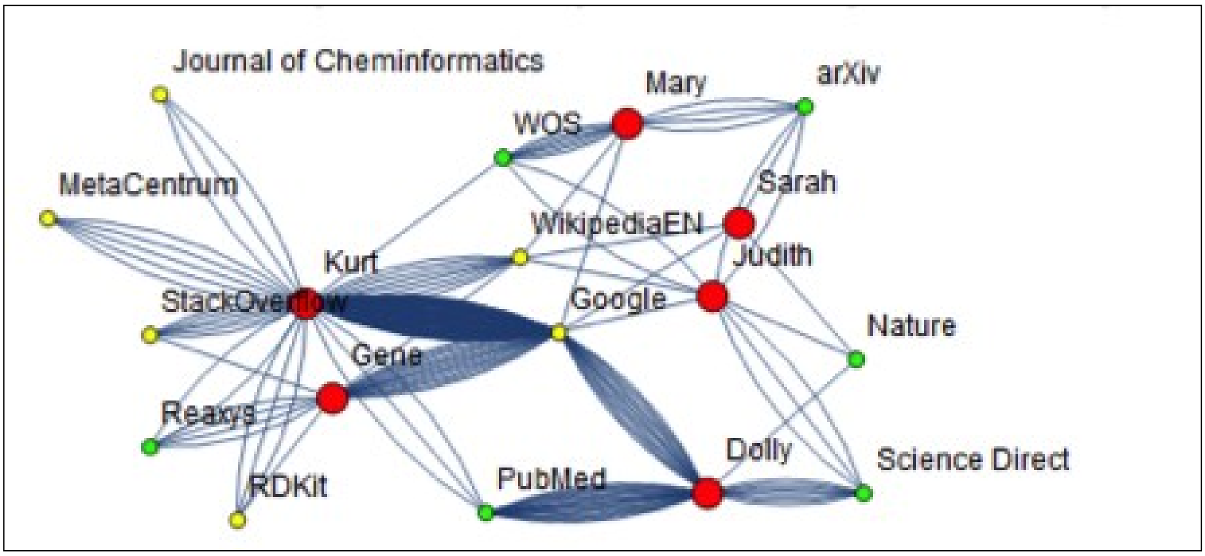


Figure 3: Visualization of Concept: Virtual Interactions as Fieldsites (for detail, see Section 5.2.2.1 and Appendix G).

For my project, including these virtual field sites meant the study was multi-sited in both in the physical world (i.e., field sites being the workplaces of research participants) and in the virtual GNAE. I anticipated this hybrid multi-sited approach would 1) increase the depth of description for the information-related activities observed, 2) place the primary site of observation within the GNAE itself, and 3) allow me to compare and analyze the attributes of nodes (e.g., an expanded set of the central categories listed in Table 1 above).

Additionally, this kind of conceptual framework meant the research approach could be repeatable for any group of scientists. I assumed that findings using this approach would become more generalizable the greater the number of participants and the greater the number of virtual interactions observed. Unlike in IS metrics studies utilizing network science methodologies, the focus of observation here is the *virtual interaction itself*, not the "trace" of data which has been published or expressed (e.g., website usage data, bibliographic data) in harvestable format. Because of this, it opens up the possibility of observing what scientists are *doing beyond the library paywall* and what scientists are doing informally; i.e., what kind of information they work with that might not be preserved or "authenticated" by an institution in the traditional library/archives sense (Skenderija 2009)—in this way, the concept using virtual interactions as field sites I observed enabled me to answer RS1 and RS2.

From these initial research design thoughts I created my thesis proposal, submitted to and approved by my advisor in November 2012. At that point, I anticipated I would be studying only the information behaviors theoretical physicists because I had two potential contacts for possible participants in this area.

3.1.4 Limitations of Research Design

Regarding screenshots, there is a possibility that participants might self-select or represent certain information-related activities in taking screenshots, therefore exhibiting a kind of Hawthorne effect in which actual behavior is modified. While recent analysis of the original Hawthorne experiments question the existence of the effect itself (see Levitte and List 2001), it is possible research participants might not represent all of their information-related activities in screenshots. However, the intent of this study is not to be comprehensive in scope but is rather to initially explore patterns illustrated by the data at hand. I would argue that self-selected screenshots are no more biased than survey responses, which can also be modified in relation to the questions at hand. I believe having multiple participants in the study, in different locations and in different research areas, combined with a comparison against artifacts gathered during conventional fieldwork as was the case in this study does reduce such possible effects by providing a broader spectrum of data sources than with fewer participants, but such an effect cannot be fully excluded.

Howard (2002, p. 554) cautions online ethnographers to avoid organizational and technological determinism when designing virtual ethnographies, reminding researchers to carefully consider their analytical frames in relation to fieldsites selected for a project. I believe the multi-sited framework with its essential comparative analytical posture does counteract possible risks regarding organizational and technological determinism, as do reflexive descriptions of analysis and interpretive processes.

As Pepe (2012, p. 2) notes citing Stodden (2009), "data reusability and reproducibility of results are crucial to modern scientific communication." I designed this project so the multi-sited approach might be used/re-used over similar or different populations, but results of future research projects would of course vary according to the populations under observation.

3.2 Logistical Preparations for Fieldwork

3.2.1 Summary

With my research design in place and initial research proposal approved, I made preparations for entering the field by reading additional anthropological descriptions of fieldwork methods and considering how I might reach out to several actively-working scientists with whom I could discuss my initial research design thoughts and procedures as a pilot study. At this point, I was not yet enrolled as a student and Humboldt did not yet have an Institutional Research Board or other similar entity, so I did not provide consent forms to my pilot discussion partners. I did, however, make all of them aware via email that our discussions would be used for my future doctoral research project and had—even at this early stage—acquainted myself with the Statement of Ethics of the American Anthropological Association (AAA 2012).

As such, throughout this dissertation, I will not refer to my research participants by their real names or by their specific institution in order to protect their privacy. In providing research interests, I have also generalized the level of scientific specialties (e.g., for Participant 1, condensed matter physics without sub-specialty) to avoid identification of participants via Google. Official participants (those who signed research consent forms) were assigned, in chronological order of participation, numbers starting with 1 (p1, p2, etc.). I refer to participants starting in Chapter Four with pseudonyms in order to make the narrative less clinical in tone. Unofficial participants (those who did not return consent forms to me but with whom I had some kind of research contact) received numbers prefixed by *n* (n1, n2, etc.; "non-official participants"). While I do not identify institutions at which participants conduct their work by name, but I have included Carnegie Classification information for institutions in the United States (Carnegie Foundation 2010). Institutions outside the US were, with the exception of one research institute, large public research universities. Additionally, I have included geographical identifiers at a level of granularity which I feel does not comprise the privacy of my research participants (country, US state—or continent, in the case of one potential participant introduced in Section 3.2.2).

3.2.2 Pilot Emails, Discussions, and Observations

In October-November 2012, I made contact with two theoretical physicists via email—the first, a friend of a friend and Assistant Professor at a private RU/H: Research University (high research activity) in New York State; the second, a high-energy theorist at a well-known European research institute of whom I had been made aware by a future colleague in the Czech Republic.

In both cases, I provided these two contacts with my initial research proposal, to gauge if my fledgling thoughts about data design, collection, and overall emphasis made sense to them, and to ask if they might know potential research participants. Intuitively, I realized “entering the field” would be difficult, and was—even in 2012—concerned about reaching minimally $n>1$ research participants. If these two physicists might become active participants, they might assist me in finding more participants on at least two continents.

One of these initial contacts became a key informant during the first two years of the research project, and the other was the first to establish a pattern I experienced throughout the study when reaching out to possible participants: he initially said yes, he could help—and upon my providing additional information about the project via email, he went silent. A summary of my efforts in reaching out to potential participants is found in Section 3.3.

In preparation for these initial discussions, I read about the scientific background of both contacts, to at least acquaint myself with their research interests and the vocabularies of their disciplines. For my key informant, for example, I read the following Wikipedia pages and asked her about her involvement (if any) in their editing: string theory, soft matter, soft materials, granular material, amorphous solid (S. Krueger, 2012, pers. comm. p1, 6 Nov).

The second contact (unofficial participant n1), in the only email I received from him, appeared to think that studying theoretical physicists might not be the best subjects for studying collaboration patterns—which, however, was not the primary focus of my study. He noted that “on the theory side, articles and projects are often run by a handful of individuals, most like in other branches of physics” (2012, pers. comm. n2, 1 Nov).

In January 2013, I made plans for a set of in-person discussions and observations at the workplace of the Assistant Professor I had emailed in November, as I would be living near her workplace for one month prior to my departure for Europe in March 2013. I contacted, in addition to the Assistant Professor, the managers of a well-known science database (n2 and n4, at an RU/VH: Research University with very high research activity, United States) and asked for an appointment to learn more about the database’s administration and day-to-day management; n2 provided me with email contact information for another theoretical physicist (n3 at an RU/VH). Thus, I emailed n3, and through him became aware of the aforementioned MMR work by Velden and Lagoze (2013). He was open to assisting me with specific questions via email. I did not approach him for assistance in locating future research participants because I feared he would be too busy to assist me in the project.

In February 2013, I met with in-person with n2 and n4, learning more about their database, which is heavily used by scientists and which has a global scientific advisory board. This conversation provided me with background information about the scholarly publication practices of scientists which I felt would help me to better understand the research habits of future project participants.

Finally, I conducted a “pilot field observation” with p1 for one month at one of her worksites. During this period, I took a set of field notes which I shared with her to check for accuracy, discussed my research plans with her (including possible data-gathering methods), and casually observed her at work—she would typically conduct her research at her laptop late at night on a desk at a kitchen table or in a nearby “office cabin” built for this purpose. Participant 1 would also Skype from the same locations with her collaborators. The Skype conversations I observed were about relationship building (e.g., discussing family matters) rather than about sorting out research details, which appeared to have been handled primarily over email or phone instead of videoconference. According to my fieldnotes (S.Krueger 2013, Fieldwork Transcription p1, 27 Feb):

> [P1] uses the phone occasionally for collaboration, while driving and with specific question (e.g., about an equation)–particularly if she knows the collaborator well.

While observation was helpful in building my own relationship with p1, it was not very useful in terms of learning about information-related behaviors themselves. Conversations (informal interviews) and field notes yielded better detail about such behaviors; for example, in one discussion captured by field notes, p1 told me Web of Science is "an important resource" and how she "was working with a colleague using Surface Evolver [a program for modeling liquid surfaces, Brakke 2013]" (S.Krueger 2013, Fieldwork Transcription p1, 27 Feb). With this information, I was already able to test-plot codes against these resources, even without knowing more about p1's exploration and discovery behaviors.

Table 3: Pilot Coding Test for Participant One.

Resource	*RQ: Codes*	*Behaviors and Notes*
Web of Science	RQ1 and 2: LIB, FORMAL Additional: COMM	Important commercial resource provided by library infrastructure
Surface Evolver	RQ1 and 2: NON-LIB, INFORMAL, Additional: NON-COMM	Software for modeling liquid surfaces according to forces and constraints; non-library, non-profit

Even during the pilot, the importance of having *bounded* my project to specific research questions in IS became apparent: while contextual information about a particular participant's actual scientific work proved useful for establishing rapport with contacts, it was clear I would need to gather more data about information-related behaviors going forward rather than focus on the details of scientific practice.

3.2.2.1 Summary of Pilot

In sum, this pilot made me feel more certain that virtual observation, though perhaps difficult to achieve, would be crucial to answering RQ1 and RQ2. Physical observations and discussions, while helpful and essential for triangulation purposes, would simply add context to the information-related behaviors captured—frozen in time, so to speak—in screenshot (originally conceived of as video) form. To supplement visual content, I began considering the necessity of what would be my future pre-interview questionnaire (Appendix D)—a tool which would standardize my initial questions about information-related behaviors with participants which I could analyze/code and compare to other questionnaire responses and to future visual data.

Reaching out to and building relationships with pilot participants took time, particularly in terms of logistical issues (e.g., planning meeting times, awaiting email responses to questions). I therefore learned I would need to be patient and persistent in attempting to find future research participants. Participant 1 and I discussed possible strategies for dealing with this; she would at some point try to help me locate participants, if needed; she additionally noted “summer would be a good time to conduct the actual research & video capture observations, when people are not teaching” (S.Krueger 2013, Fieldwork Transcriptions, p.2, 27 Feb).

Finally, I discovered during the pilot that demographics and other background information about potential participants could easily be found online for actively-working scientists, who usually have public profile pages at the institutions where they work and which are often supplemented by profiles on LinkedIn, Google Scholar, Mendeley, and Research Gate. This meant I would not need to explore such questions in my pre-interview questionnaire or in initial discussions with participants.

Table 4: Pilot Study Participants Summary.

Research ID (n=no consent form; email consent only)	*Gender*	*Age Range*	*Institution Type*	*Public or Private*	*Subject*	*Research Interests*	*Location*	*Contact Via*
1	F	40-50	RU/H	Private	Theoretical Physics	Condensed Matter Theory	NY, US	Friend of a friend
n1	M	30-50	Research Institute	Public	High-Energy Physics	Physics Constants	Europe	Czech Colleague Recommendation
n2	M	30-50	RU/VH	Private	Scholarly Publishing	XML Applications	NY, US	Random US (my choice)
n3	M	50+	RU/VH	Private	Theoretical Physics	Quantum Field Theory	NY, US	US Colleague Recommendation
n4	M	30-50	RU/VH	Private	Scholarly Publishing	n/a	NY, US	Random US (my choice)

3.2.3 “Embedding” in Central Europe and Standardizing Processes

3.2.3.1 A Czech Science and Technology Campus

In March 2013, I moved to the Czech Republic and began working at a science and engineering library (NTK) in the heart of an engineering campus. This meant I became embedded as a participant observer in physical and virtual environments surrounded by actively-working scientists and their students. The project officially became multi-sited on two continents.

NTK and its campus partner institutions were founded in the early eighteenth century by the Hapsburg Monarchy and its constituent Kingdom of Bohemia; the focus of both the original engineering school and library was military engineering support for the monarchy

(NTK, 2015). Over the past three hundred years, both the library and the universities have changed their names and locations on several occasions, and the library was separated from the academic institutions following World War II. In 2009, a new library building opened within a short distance from the former institution it had originally served and which is now split into two independent universities, the Czech Technical University in Prague and the University of Chemistry and Technology, Prague. NTK provides library services to the latter (ChemTK). The Technical University has its own separate library also housed in the same building as NTK.

Figure 4: Campus Dejvice with NTK as Number 12 (CTU 2011).

Being so embedded, I was able to informally observe the information behaviors students and professors on a daily basis—not only in the physical realm on campus, in meetings and in casual conversations, but also virtually in relation to library-related information activities online.

While I made the fact that I was conducting doctoral research public on my LinkedIn and library profiles, I will not include data gathered via informal observation, discussions, or library statistics in this dissertation *except* when it is related or relevant to work with my formal research participants, all of whom signed informed consent forms. In other words, I used "my own backyard" (Creswell 2014, p. 188) only in order to provide me with contextual background and to test services developed in response to observations during the research process. I describe several services in Section 5.2.2.

3.2.3.2 Standardized Participant Paperwork and Process

As a result of the pilot study, I realized by summer 2013 that I needed a packet of standardized materials which would assist me when reaching out to potential project participants, as well as a standardized process which I could customize with minimal effort. While I modified some wording of individual documents over time, the basic process has proven helpful to date and would be reusable in future research projects:

1. Receive potential participant contact email from friend or colleague
2. Email participant, noting who let me know about potential participant, attaching most recent version of thesis proposal (later: attaching Informed Consent Form [Appendix B] and Request for Participation letter [Appendix C]). In email, highlight key deliverables and approximate time investment required during the research process for the potential participant minimally:
    - Pre-Interview Questionnaire (Appendix D)
    - Video Screenshot Capture (later: screenshots)
    - Follow-ups.
3. Potential participant a) says yes, b) declines, or c) become silent.
4. For participants who say yes (later: those who send back Informed Consent Form), send them the Pre-Interview Questionnaire.
5. For participants who return Pre-Interview Questionnaire, a) ask by email or in person about any follow-up questions regarding questionnaire and b) ask for video/screenshot captures (screenshots for a two weeks, if possible).
6. If participants respond at this point, work more closely with them to get any and all follow-up contextual data and develop a long-term rapport.

This framework functioned smoothly throughout the study, although acquiring formal participants remained the most difficult and time-consuming component of the project, as will be seen in Section 3.3.

I concurrently conceived the concept of an ideal "full research cycle" for each formal participant, which would minimally include: 1) completion of the Pre-Interview

Questionnaire, 2) at least one interview/discussion (via email, Skype, or in-person), and 3) Screenshot Data (two weeks, if possible).

3.2.4 Formal Approval

In Spring 2013, I formally enrolled as a doctoral student, which enabled me to apply for research board approval of my project. Upon recommendation of my primary advisor, I created a Request for Participation Form and a Consent Form (Appendices C and B) to submit to the head of the Humboldt Institute for European Cultural Anthropology's Examination and Ethics Committee, since (at that time) Humboldt did not have a university-wide IRB.

The project was approved without any objections on July 7, 2013; Dr. Seadle received formal notification of this from Prof. Dr. Stefan Beck (now deceased; Appendix H). This formal approval letter certified that my study followed the guidelines of the Code of Ethics of the American Anthropological Association (AAA 2009).

With a process for contacting potential research participants in place together with formal approval, I then felt confident reaching out to potential participants. At that time (spring 2013), both my key informant p1 and I were fairly confident we would be able to find a core set of formal participants with not too much effort. In this, both of us were very much mistaken.

3.3 Entering the Conventional Field: A Time-Space Odyssey

3.3.1 Summary

Finding a substantive (n>1) number of formal participants for the study became an ongoing process throughout the entire project. Concentrated efforts to recruit potential participants took place in the summers of 2013 and 2014, with communication continuing with active participants in the interim periods.

3.3.2 Summer 2013

In Summer 2013, key informant p1 and I made a concentrated effort to recruit theoretical physicists for the study. Our plan was to utilize her prior article collaboration networks; she would send the initial contact email (Appendix E), and I would follow-up following the steps listed above. This group was spread across two countries (US and UK). Participant 1 noted (p1 2013, pers. comm. 17 July):

> These are people that I know and have collaborated with. They may still not respond, but perhaps they can forward it to others, which can then lead to some responses—the network effect.

These efforts were, however, unsuccessful, resulting in only one email response—"I would be happy to participate in the study" (n5 2013, pers. comm. 8 August 2013). After I sent n5 the Informed Consent Form, she became silent.

Key information p1 remained open to helping find participants in the study (p1 2013, pers. comm., 23 July):

> Unfortunately, I can't strong arm anyone into participating, but let's see what happens. I may also put down that person's participation can be used as an outreach item in research proposals that also require some form of outreach. I should have thought of that earlier. That is good incentive. Also, I did just meet a scientist named [x]. I will be meeting him in person at Princeton in early September. He is a scientist interested in networks. He works on a lot of different things. Given his research program, one would think he would be very interested in participating. When I meet him in September, I will mention your thesis work to him and feel him out about it.

In parallel to the theoretical physicist campaign, I made contact randomly, at a semester-long Czech class, to an experimental physicist conducting post-doctoral research in condensed matter physics in Prague; she (p2) became a formal participant. Through her, I was able to recruit one more research participant (p3). I was able to complete the entire research cycle with p2 (completion of the pre-interview questionnaire, in-person observation, interview follow-up, screenshots) and with p3, I was able to complete the data-gathering cycle except for screenshots (completed pre-interview questionnaire, in-person observation, follow-up) before her interest in the project waned.

Table 5: Participant Outreach, Summer 2013.

Research ID (n=no response or consent form not returned)	*Gender*	*Age Range*	*Institution Type*	*Public or Private*	*Subject*	*Research Interests*	*Location*	*Contact Via*
p2	F	20-30	Research University	Public	Condensed Matter Physics	Cystallography	Prague, Czech Republic	Random, Czech Republic
p3	F	20-30	Research University	Public	Condensed Matter Physics	Quantum theory of condensed matter	Prague, Czech Republic	Czech Participant Recommendation
n5	F	30-50	Master's L	Private	Elasticity and statistical mechanics	Cell mechanics	NY, US	US Participant Recommendation
n6	M	30-50	n/a	Public	Biophysics	Soft condensed matter	CA, US	US Participant Recommendation
n7	M	30-50	RU/VH	Public	Polymer Science and Engineering	Statistical mechanics	MA, US	US Participant Recommendation
n8	M	30-50	Research University	Public	Soft condensed matter	Biological and artificial membranes	United Kingdom	US Participant Recommendation
n9	F	30-50	Research University	Public	Condensed matter physics	Active matter	United Kingdom	US Participant Recommendation

3.3.3 Between Summers (2013-2014)

P1 tried again, in November 2013, to ask her collaborator n5 about participation during an in-person meeting, without success.

In December 2013, I discussed the project with a colleague from the Czech Academy of Sciences, and he felt confident he had a promising lead to a potential participant. However, I received no response from this contact after emailing him.

Anticipating the summer window of potential research opportunity, in May 2014, I asked a work colleague from a partner institution about possible local participants. He provided me with two names, both of whom were doctoral students actively conducting research. I emailed them and they became my second and third key informants (p4 and p5), immediately returning consent forms and pre-interview questionnaires to me.

My primary thesis advisor also provided me with an email introduction to two chemists, who initially said they might participate but who then failed to respond to post-introduction emails.

As seen by the fact that my Czech colleagues assisted me in this process, my “embedding” myself in the field and developing professional working relationships with scientists beyond this study provided me with the ability to gain additional research contacts, two of whom became crucial study participants. This was similar to my experience with p1, with whom I had spent one month during the pilot study.

Table 6: Participant Outreach, Between Summers (2013-2014).

Research ID (n=no consent form; email consent only)	*Gender*	*Age Range*	*Institution Type*	*Public or Private*	*Subject*	*Location*	*Contact Via*
p4	M	20-30	Research university	Public	Cheminformatics	Prague	Czech Colleague Recommendation
p5	M	20-30	Research university	Public	Cheminformatics	Prague	Czech Colleague Recommendation
n10	M	50+	Research institute	Public	QCD	Prague	Czech Colleague Recommendation
n11	M	50+	n/a	n/a	Analytical chemistry	NY, US	Advisor Recommendation
n12	F	50+	Master's L	Public	Technetium (Tc)	NY, US	Advisor Recommendation

3.3.4 Summer 2014 to Present

I held my first research-related in-person meeting with p4 and p5 (both whom I had actually met earlier at several informal work-related gatherings) in mid-June 2014, in their shared office, a small narrow space with two desktop computers in a building near NTK. This space is periodically filled with computer equipment these participants use for instructional purposes (Figure 4). I have, since then, continued to meet regularly with these participants in order to better understand their research and to clarify questions about data I have gathered. Both p4 and p5 have completed the entire research cycle (pre-interview questionnaire, screenshots, observation, follow-up discussions).

Figure 5: Teaching Computers in Workspace, p4 and p5 (March 2015).

In August 2014, I also reached out to a friend of a friend in the US, a microbiologist; this potential participant declined because of work-life issues: “My life has been a bit of a whirlwhind [sic.] of late and I'm no pro at time management to begin with. So not sure it'd be fair to you to promise you 10-15 hrs” (n13 2014, pers. comm., 23 August).

In late 2014, I also mentioned my need for participants at a Humboldt doctoral seminar, which resulted in one additional participant (p6). She has, like p1, 2, 4, and 5, completed the entire core research cycle.

During this period, I considered asking several other professors with whom I collaborate closely to participate in the study, but decided against it because I felt it might be overstepping the boundaries of our working relationship in another context.

Table 7: Participant Outreach, Summer 2014 to Present.

Research ID (n=no consent form; email consent only)	*Gender*	*Age Range*	*Institution Type*	*Public or Private*	*Subject*	*Research Interests*	*Location*	*Contact Via*
6	F	40-50	RU/VH	Public	Toxicology	Environmental and clinical	MT, US	Friend of fellow doctoral student
n13	M	30-40	Master's L	Public	Microbiology and Immunology	Virus-cell interactions	LA, US	Friend of friend

3.3.5 Summary of Conventional Field Entry Process

I conducted a pilot and more formalized fieldwork physically in two locations: New York (US) and Prague, Czech Republic, with one month of participation observation in the first location and since March 2013 at the latter location (NTK and the Prague-Dejvice technology campus). In 2015, I added a virtual site (Montana, US) to the study.

Altogether, six formal participants, four women and two men representing four different research universities (three public and one private) completed significant parts of the research cycle, with five participants providing vital screenshot data to me. Two of the women are ranked professors (Assistant and Associate Professors). Two women were postdocs during their participation in the study; one is now a professor at the institution where she was a postdoc and the other was conducting research in another field in Germany as of my last contact with her. The two men are doctoral students actively conducting cheminformatics research. All formal participants are under fifty years of age, with four participants being under thirty at the time they started participation in the project. I interacted with all participants in English. Two participants are native English speakers, Czech is the mother tongue of three participants, and one participant is a Russian native speaker.

I was connected to all but one participant by colleagues or friends (i.e., in the loosely-defined sense of “snowballing” as in Walsh 2014).

Table 8: Formal Participant Summary (n=6).

Participant	*Subject*	*(State or City), Country*	*Native Language*	*Connection*
1 (key informant)	Theoretical Physics	New York, US	US English	Friend of Friend US
2 (key informant)	Condensed Matter Physics	Prague, Czech Republic (CZ)	Russian	Random CZ
3	Condensed Matter Physics	Prague, CZ	Czech	Colleague of Participant CZ
4 (key informant)	Cheminformatics	Prague, CZ	Czech	Colleague Recommendation CZ
5 (key informant)	Cheminformatics	Prague, CZ	Czech	Colleague Recommendation CZ
6	Immunology & Infectious Diseases	Montana, US	US English	Friend of Fellow Doctoral Student US

Summing up non-participant contact, I attempted to gain additional participation from a European nation (country not disclosed for privacy reasons), the United Kingdom (two locations), the Czech Republic, and four US states (New York, California, Massachusetts, Louisiana). This group included ten men and three women and all were over thirty years of age. While these potential participants did not complete the research cycle, I was able to add depth to my understanding of their practices by reviewing their online profiles, research interest information, and article publishing threads.

Table 9: Potential Participant Summary (n=13).

Participant	*Subject*	*(State or City), Country*	*Connection*
n1	High-Energy Physics	Europe	Friend of Friend US
n2	Scholarly Publishing	New York, US	Random US (my choice)
n3	Theoretical Physics	NY, US	US Colleague Recommendation
n4	Scholarly Publishing	NY, US	US Colleague Recommendation
n5	Elasticity and Statistical Mechanics	NY, US	US Participant Recommendation

n6	Biophysics	CA, US	US Participant Recommendation
Participant	*Subject*	*(State or City), Country*	*Connection*
n7	Polymer Science and Engineering	MA, US	US Participant Recommendation
n8	Soft Condensed Matter	United Kingdom	US Participant Recommendation
n9	Condensed Matter Physics	United Kingdom	US Participant Recommendation
n10	QCD	Prague, CZ	Czech Colleague Recommendation
n11	Analytical Chemistry	NY, US	Advisor Recommendation
n12	Technetium (Tc)	NY, US	Advisor Recommendation
n13	Microbiology and Immunology	LA, US	Friend of Friend US

As noted above, finding and keeping formal research participants, defined as those who completed and returned an Informed Consent Form, was a time-consuming and labor-intensive process which involved concentrated relationship building (over email or in-person). Because none of the non-participants stated reasons for why they did not wish to participate in the project, I have no data explaining what might make participation more likely in future studies. I feel, however, based on my observations in the field, that time and publishing pressures play a large role in the difficulty of recruiting actively-working scientists into an ethnographic study. For example, key informant p1—author of over thirty publications in high-ranked journals—often worked and wrote in the middle of the night as her family slept; she viewed participation in the study as a form of mentorship.

I believe receiving the backing of a professional society or the leadership of an institution under study would potentially improve participation rates (i.e., as was the case in Allard et al. 2007 with IEEE; also with Upadhya 2008, who worked with top managers at IT companies in India to gain access to research subjects). However, such organizational backing might make participants less reluctant to share data—particularly screenshot data. Adding an organizational partner into the mix might potentially bring up theoretical issues regarding “power, resistance and subjectivity” (Upadhya 2008, p. 68) in relation to a sponsoring organization. Being a “lone, independent researcher” provided me with a level of trust and engagement with formal participants without introducing such power and subjectivity issues into this project.

3.4 Entering the Virtual Field

3.4.1 Summary

When considering the virtual field, double articulation in the sense of Deleuze and Gauttari 1987, p. 40) comes to mind, in which phenomena manifest themselves as pairs ("the Lobster, or a double pincer, a double bind"), with content and expression intermingling, multiplying, and dividing "ad infinitum" (p. 44). As seen in Chapter Two (and in the data analysis and interpretation sections), a multi-sited ethnography exhibits such an intertwined duality: within the GNAE, the work of scientists is both local/global, physical/virtual—these double articulations combine and complement each other. The necessary virtual field complements the conventional setting, and vice versa.

The virtual field itself is not just one entity (refer to the definitions of GNAE in Chapter Two); it is the abstract plane of articulation for endless virtual interactions between many networked components. My key task in the virtual field was to find points of interaction of scientists with networked information which I could observe and analyze in relation to RQ1 and RQ2.

For contextual background, I have been monitoring online statistics for NTK and its partner, ChemTK, since March 2013 to date. These statistics give me insight into what is happening within the framework of library-supported resources, but provide no help in observing what is happening beyond the library paywall—i.e., what scholars are doing beyond library-supported IP ranges, proxy servers, and web pages which include Google Analytics' tracking code.

To get beyond the paywall, I therefore asked formal participants about their information-related behaviors (in the Pre-Interview Questionnaire and in discussions with them) and supplemented/compared this with screenshots they provided (i.e., emic expressions of their own information behaviors), supplemented in turn by observations of working environments in the physical world (participants 1-5) and online. I will provide detailed information about data gathered for each participant in Section 4.6.

Table 10: Physical and Virtual Duality (Key Informants p1, 2, 4, 5).

Physical			*Virtual*		
Participant	*Kind*	*Duration*	*Participant*	*Kind*	*Number/Duration*
1	On-Site observation and Skype interviews	One month (informal), in-person interview 1 hour	1	Screenshots	30/August 2014, two weeks
2	In-person conversations	Five hours	2	Screenshots	3/unsure
4	On-site observation and discussions	Six hours	4	Screenshots	112/August 2014-January 2015
5	On-site observation and discussions	Six hours	5	Screenshots	18/August 2014-February 2015

Participant 3 did not provide screenshot data and I only had virtual interviews with participant 6 (see Table 12 for details).

3.4.2 Initial Concept: Video

Many virtual ethnographies involving the communicative and collaborative behaviors online require the observer to become an active participant in discussion forums and other spaces which facilitate communication between members (e.g., Hine 2000). Because my primary research questions were related to virtual interactions within the GNAE, however, I felt that video screen captures would provide access to details of interaction at the level of granularity I required for this study. As Heath and Hindmarsh note in *Analysing Interaction* (2002, p. 8):

> [Video recordings] allow us to capture versions of conduct and interaction in everyday settings and subject them to repeated scrutiny using slow motion facilities and the like. Thus, they provide access to the fine details of conduct.

At the outset of this project, I naively envisioned my participants eagerly embracing the concept of video screen captures. I identified and tested a free tool (TechSmith's Jing), worried about where I might store large video files and when I would find time to transcribe video data, and pitched the idea of using video to my key informants at various times. The result was resounding silence; all participants simply ignored my questions about video research and did not comment upon it in email or in person, and I did not feel comfortable pushing them towards a response. I believe this has a lot to do with privacy concerns, but have not been able to date to broach this topic significantly with any of my participants. As a result, I switched my strategy in 2014 and came up with the idea of having participants take screenshots of what they defined as research behaviors—ideally for a two-week period.

I consider this change of tactics as an example of the difficulty of entering the virtual field at the point of interaction. Getting useful data of this kind, however, is not impossible, as will be seen in the next section. I view the fact that my key participants, by providing me with still screenshots, overcame whatever their concerns were about video—indeed, that they even provided me with screenshots at all—to be an indicator that my efforts in the conventional field to establish trust paid off. The screenshots comprised essential *gifts* from my key participants and they enabled me to answer my primary research questions.

3.4.3 Compromise with Rich Results: Screenshots

The screenshots given to me by participants enabled me to “see there” and to enter their information lifeworlds. I consider these images to be still video frames which freezes virtual interactions in time (Spinney 2011 provides additional background on theories of video use in ethnography, including mobile video and freeze-frame). My background in academic information resources coupled with the knowledge I gained in the field provided me with the ability to deeply describe what was happening at points of virtual interaction with the GNAE. I could also graph these data in various ways to complement and illustrate textual narratives and to identify patterns for interpretation.

While these data represent a compromised version of my original video vision, in they were ultimately the most useful artifacts in this study for addressing RQ1 and RQ2.

3.4.4 Supplementary Virtual Information

Virtual fieldwork in a more conventional sense involves research into the lives of participants (and non-participants) as represented online (e.g., mining demographics and publication information as well as information about collaborating research partners)—in other words, learning more about participants’ information lifeworlds.

Supplementary online research enabled me to better-understand the professional careers of my participants, to review terminology I did not understand, and to conduct research about the providers of resources identified in the screenshot data (such for- or non-profit status and country of origin). Such contextual information is readily available online.

3.4.5 Summary: Entering the Virtual Field

Without conducting conventional fieldwork and establishing trust relationships with my key informants, I would have been unable to sufficiently enter the virtual field in relation to my primary research questions. While I had to abandon my original concept of gathering video screencapture data, I found the resulting still screenshot data to be insightful and *deep enough* for the purposes of this dissertation. As will be seen in Chapters Four and Five, these representations of visual interactions were essential to this project. Virtual fieldwork additionally strengthened the multi-sited nature of the study, providing essential, reciprocal (in relation to conventional fieldwork) duality for analysis, comparison, and interpretation.

3.5 Summary of Procedures and Research Methods

This project was conceived as an exploratory one which could be repeated in and extended to different settings and populations. As such, it provides a model for future researchers and the model could be extended to include quantitative components for statistically-random populations, which was not the case in this study for reasons identified in Section 3.3. For me, actively-working scientists were a hard-to-reach (although not impossible to reach) population, and I resorted to a lengthy process of non-random "snowballing" via colleagues and friends to recruit formal research participants.

While I had originally envisioned video screenshots to be essential for entering the virtual field, still screenshots provided me with a mechanism for observation and deep description of virtual fieldsites—points of interaction of my research participants with information in the GNAE. Data gathered during the research cycle will be described in Chapter Four below and interpreted in Chapter Five.

Chapter 4 Formal Data Description

4.1 Summary

I gathered, transcribed, coded, and analyzed primary data from six participants. As expected from a multi-sited ethnographic study, types and amounts of data gathered varied according to participant and fieldsite.

As noted in Section 3.1.2, final data types differed from my original conception (see Table 2, Initial Data Sources Ideas). In the end, asynchronous data types (email, screenshots, pre-interview questionnaire, supplemental online information) were easier to gather than those collected in a synchronous manner (e.g., transcripts of live, synchronous Skype interviews).

Table 11: Summary, Data Gathered.

Data Gathering Technique or Source	*Data Type*	*Data Use*
In-Person Observation	Primary data; scratchnotes or captured in email	To supplement and enrich virtual observations and visual data (RQ1)
In-Person or Virtual Interview or Discussions	Primary data; scratchnotes or captured in email or on paper; synchronous and asynchronous; conventional and virtual	To observe participants in their own work settings, to build relationships with participants, and to clarify and enrich other data sources, particularly screenshot data (RQ1)
Skype Interview or Discussion	Primary data; asynchronous scratchnotes or captured in email representations of synchronous virtual discussion	Same as above
Email Correspondence (including pre-interview questionnaire discussion)	Primary data; emails; asynchronous, virtual	To share and discuss project details with participants at any point of the research processes
Photographs of Worksites	Primary data; image files; synchronous during conventional fieldwork	To supplement worksite descriptions from in-person observation or discussions
Screenshots	Primary data; image files (shared via email, Dropbox, or Google Drive); asynchronous, virtual delivery and description	To observe and analyze virtual interactions with information within the GNAE (RQ1)
Supplementary Online Materials	Secondary data; Web pages, blogs, Wikipedia entries, discussion forums, and others as identified by participants or in the course of my contextual research	To understand scientific concepts, to clarify questions about information sources, to understand collaborative networks, to trace online biographies and publishing patterns for participants (when available)

In this chapter, I will describe these data in detail as well as data management, coding, and analysis techniques—both at a general level and in relation to each participant in the study.

4.2 Overview of Data Gathered

My datasets evolved over the course of the project as I entered the conventional and virtual fields and extended the number of research participants over time and location. I utilized the standardized research process outlined in Section 3.2.3.2 to guide me in the data gathering process, adjusting my techniques according to the realities of fieldwork and to the level of engagement with each project participant. I believe my data gathering and management techniques improved over the two and a half year duration of the study. My experience in managing the research cycle with my most recent participant (p6) illustrates this—we were able to conduct the entire cycle virtually and with a concentrated focus on information-related behaviors and less emphasis on issues not related to my central research questions.

I had originally planned to manage data using an open source qualitative analysis system, but ended up creating my own approach to storing and manipulating my data because this gave me more flexibility in terms of considering issues and with displaying data in a manner which made sense to me. I also had an eye on any future possible open data requirements—my system would enable me to easily share coded transcripts and coded screenshot lists with others if needed in the future without compromising the privacy of my research participants.

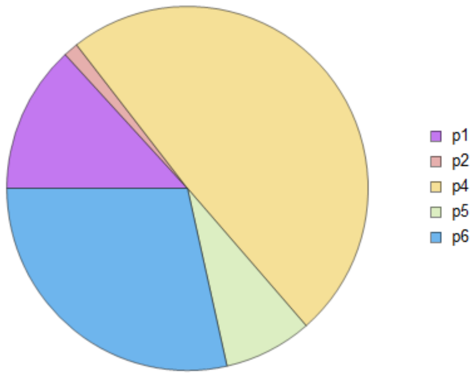
I therefore organized data into two main categories: 1) outputs of fieldwork and 2) screenshots taken by participants which I could subsequently transcribe, code, and analyze. I summarize data in these two categories in the tables below and provide in-depth analysis by participant in Section 4.6, where I also introduce pseudonyms for each participant.

Table 12: Fieldwork Outputs: Summary by Participant.

Participant	*In-Person Observation*	*In-Person Interview or Discussion*	*Emails*	*Skype Discussion*	*Total Coded Interactions*
1	One month (informal); scratchnotes captured in coded fieldnotes	One hour captured in coded fieldnotes	Sixteen substantive interactions coded in fieldnotes	Two hours captured in coded fieldnotes	20
2	Two hours; scratchnotes captured in coded fieldnotes	Two hours; scratchnotes in email form read and commented by participant	Eight; coded in fieldnotes	None (in Prague)	12
3	None	One hour; scratchnotes in email form read and commented by participant	Three	None (in Prague)	4
4	Half hour; scratchnotes captured in coded fieldnotes	7.5 hours; scratchnotes captured in coded fieldnotes	Thirteen	None (in Prague)	22
5	Half hour; scratchnotes captured in coded fieldnotes	7.5 hours; scratchnotes captured in coded fieldnotes	Two	None (in Prague)	10
6	None (in US)	None (all interaction by email to date)	Three	None to date	5
Total All Participants	**One month (informal) plus three hours**	**19 hours**	**47**	**Two hours**	**73**

Table 13: Screenshots: Summary by Participant.

Participant	*Total Screenshots*
1	30
2	3
3	None
4	112
5	18
6	65
Total All Participants	**229**



Regarding emails, I only included emails *from* participants which I deemed relevant to the study. Emails I sent are not included in these totals.

4.3 Data Preparation and Readiness

4.3.1 Summary

The tables above include data I transcribed and coded. Data I left uncoded were some ephemeral emails (i.e., those which contained no substantive information related to this study), images of worksites which contained no useful information applicable to my primary research questions), and some scratchnotes regarding research specialties and not information-related behaviors.

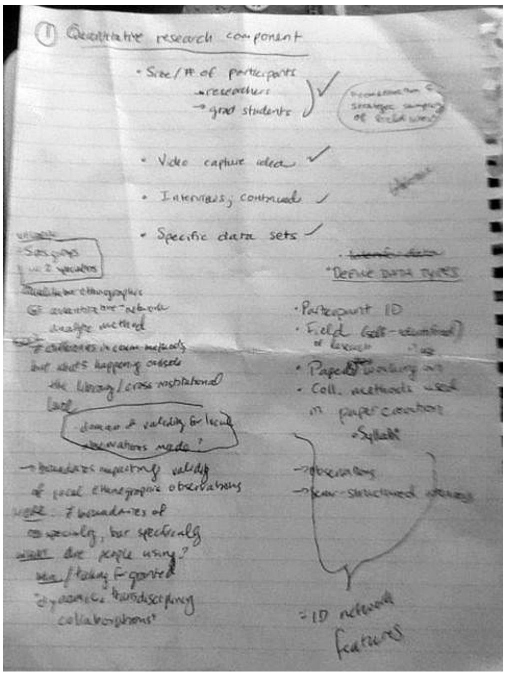


Figure 6: Sample Scratchnote.

My data preparation techniques were slightly different for fieldwork artifacts than for screenshot data, although in both cases I used a fieldnote description model similar to that described in Hartel and Thompson (2011, p. 7) citing Taylor and Swan (2005).

4.3.2 Fieldwork Data

I prepared a master fieldnotes file in Word and placed this same data in Excel files for each participant. The Word file is easy-to-read and can be shared with others. The Excel files allowed me to cross-check the accuracy of my transcriptions as well as the number of interactions for each participant and also enabled me to later interpret and graph participant-level data. The Excel versions can also be shared. Both file versions contain descriptions and locations of data from various original sources, listed by participant in reverse chronological order (oldest to most recent). I preferred this method solely because of personal preference and potential ease of sharing coded files. A database would be preferable in projects with more participants as well as those requiring multiple researchers to code and analyze data collaboratively.

These fieldnotes include artifacts gathered both in-person and virtually organized and transcribed in the following format:

- Item Number and Format
- Participant ID
- Date
- Storage Location of Original Artifact
- Theme
- Orienting Information (where applicable; notes)
- Analytic Notes (where applicable; my thoughts about an interaction)
- Analytic Commentary/CODES (see Section 4.5.2)

My fieldnotes contain no information about gestures, expressions, or behavioral tendencies for in-person observation and interviews because these behaviors were not the focus of this study. I included quotes from participants wherever possible from emails, the pre-interview questionnaire, and interviews in order to capture their emic perspectives.

Preparing the data in this manner was essential for maintaining the privacy of my participants, particularly in regard to email correspondence. Transcriptions allowed me to capture relevant information while removing details which might identify participants. I will discuss considerations regarding privacy and long-term storage of original emails in Section 4.4 below.

Table 14: Sample Fieldwork Data, Transcribed and Coded.

4. Format: Email and questionnaire and observation final review **Participant**: p3 Date: 3 Nov 2013 Storage Location: Gmail **Theme: Questionnaire and observation review, editing** Excerpt (text or image): Time-constraints: “I am a bit more busy and and [sic] answers may be delayed, howere [sic], I am interested in this project. “showed me a typical day, which begins by opening up ArXiv in the morning in order to check news and the landscape of what’s happening in the field.” LIBRARY SERVICES: “doesn’t go to faculty/departmental library but requests articles and have these sent to them.” Library catalog “doesn’t often find it useful and sometimes ends up purchasing books themselves.” Reference Manager and BibTeX (for LaTeX). “does not have a library subject specialist; typically, a more senior researcher will show students how to use research tools, how to read articles, and how to find. In the beginning this process is more difficult so the supervisor/mentor recommends and shows junior scholars what to look for. It is difficult for the research participant to imagine how a librarian might assist this process because the subject areas are so specific.” Does use older books for example Solid State Physics (Ashcrost and Mermin 1976) and Introduction to Solid State Physics [Kittel; latest edition 2005, original edition 1953] because they are well-written and the basics don’t change. Occasional collaboration with Skype and email Analytic Commentary/CODES: **ArXiv** **American Physical Society Physics Portal (Physics APS**): at least once a month **Nature:** at least once a month **Science:** at least once a month **Google (?version):** important source of finding articles and tracking citations **Web of Knowledge:** important source of finding articles and tracking citations. Sometimes delays and things might be several months behind. **Reference Manager** **BibTeX (for LaTeX)** **Review of Modern Physics, APS:** “most impact factor” **Google Scholar:** “doesn’t often use”/”too many irrelevant results”

4.3.3 Screenshot Data

I followed a similar process in preparing screenshot data as for the textual materials, including preparation of one master Word file as well as spreadsheets for each participant, all of which are sharable to the public.

I used an almost-identical descriptive format:

Item Number and Filename or other (usually date)

- Format
- Participant ID
- Storage Location
- Theme
- Orienting Information (where applicable; helpful notes)
- Analytic Notes (where applicable; my thoughts about an interaction)
- Analytic Commentary/CODES (see Section 4.5.2 below)

I considered including original screenshots in the master Word file list, but decided it would be difficult to address privacy concerns because some participant screenshots include personal data (e.g., open browser tabs in which a Gmail name is visible). I have cropped and included some screenshots to illustrate narrative later in this dissertation.

Table 15: Sample Screenshot Data, Transcribed and Coded.

97. Filename or other: 11.3.2014.1 Format: png GDrive folder **Participant**: p4 Storage Location: GDrive, local copy Theme: **Cocrystal** **Emic:** I wasn't sure what 'cocrystal' is, so I spent some time on wikipedia repairing this hole in my education. It probably came up in the context of protein--ligand interaction modelling. Analytic Commentary/CODES: **WIKI EN** **NON-COMM** **NON-LIB** **VIEW ARTICLE** **FORM** **NON-AUTH**

4.3.4 Information about Participants and Contacts

I also created spreadsheets for both participants and non-participants which included the following data. I kept this data separate from the other spreadsheets to ensure the privacy of those mentioned. I also included checkpoints for steps regarding progression through the research cycle to ensure I followed a consistent procedure with each participant. I additionally included information about each participant's institution as well as links for future reference to biographical information about each individual.

Table 16: Master Participant File Data.

• Name • Number (identifier key) • Gender • Age Range (20-30, 30-50, or 50+) • Title • Institution • Institution Type/Carnegie Data • School or Institution • Department1 • Department2 • URL (public information about participant) • URL2 (public information about participant) • Subject • Research Interests • Theorist or Experimentalist? • Institution Location (US State) • Institution Location (Country) • Said Yes Via Email • Informed Consent Form • Pre-Interview Questionnaire • Screenshots • On-Site Observation • Skype Interviews • Email Correspondence • Publication Record Review • In-Person Interview or Discussion • Connection (through colleague or friend, etc.)

I then created a sharable, anonymous participant file from which all identifying personal information was removed.

4.3.5 Supplementary Information

Additional secondary data I gathered online included web pages, blogs, Wikipedia entries, discussion forums, and others as identified by participants or in the course of my contextual research. I used these materials to understand scientific concepts, to clarify questions about information resources, to understand collaborative networks, and to trace online biographies and publishing patterns for participants (when available). I also used secondary data to find background on information resources, which I will discuss in Section 4.5 in relation to coding data.

Some of these materials reveal information about project participants, particularly information about research specialties, research groups, and article citation traces.

Therefore, I had to select a level of granularity to present in the dissertation narrative which would not endanger the anonymity of participants.

I did not summarize these secondary data in representative files; I simply kept informal lists of useful items and printed out materials I found particularly useful for future reference.

4.3.6 Master Information Resource Spreadsheets

I additionally required tracking sheets for all information providers I identified in the analysis and coding process, one for fieldwork data and the other for screenshots. This file allowed me to store information about each provider, including URL(s) and other contextual information—including who determines authenticity for the resource. The concept of authenticity will be described in detail in Chapter Five.

4.4 Data Storage and Protection

4.4.1 Summary

Following the data preparation process, I therefore had two sets of materials: 1) original data artifacts and 2) representations of fieldwork and screenshot data in condensed formats (Word and Excel files) which could be made public, coded, and interpreted.

4.4.2 Storage and Protection of Original Data Files

There are no privacy issues related to the small set of still-images from worksites, two of which are included in this dissertation above.

My scratchnotes in some cases do include information which might identify participants; these I store in their original paper format in a locked cabinet in my NTK office.

During the process of drafting this dissertation, I stored emails and screenshots in their original locations (Gmail, mail.techlib.cz, Google Drive, and Dropbox). I downloaded screenshots provided to me in the cloud to a local file on my NTK computer. Following completion and defense of this dissertation, I will print original screenshots, ask participants to delete online folders, delete the data from my local computer, and store one local copy of original digital images on one flash drive which I will store together with my scratchnotes.

I have printed relevant email correspondence to aid in the coding process, and following completion and defense of this dissertation, I will delete research emails from

Gmail and mail.techlib.cz and store only print copies of email correspondence together with the aforementioned original data files.

The same holds true for my master participation file and secondary data printouts; I will store important electronic files on the master project flashdrive and place one archival print copy in a file together with all other project materials.

4.5 Data Analysis Procedures

4.5.1 Summary

In approaching my primary data and the descriptive representations thereof, I began by conducting open coding, starting from my initial list of coding categories in relation to RQ1 and RQ2 (Table 1; see Cairns and Cox 2008, p. 141 for a detailed description of qualitative coding practices). These initial categories allowed me to examine information behaviors and virtual interactions and *to identify activities and/or resources intermediated by traditional library infrastructures (e.g., subscription databases to which access is provided by a participant's library) versus other entities such as Google or research groups (RQ1)*. I also wanted to examine characteristics of the institutions supplying information such as their for- or non-profit status and the formats of the information with which participants interacted, which I hoped might reveal something about the authenticity or trust assigned to the information resources by participants in relation to RQ1's *non-institutional, informal information* exploration and discovery component; these concepts will be discussed in Chapter Five.

As my analysis progressed, I refined my coding procedures and created the coding tables below which might assist future researchers conducting similar studies.

4.5.2 Developing Codes

I provide here a reflexive description of how I developed codes since, as mentioned in Section 3.1.1, I was unable to locate relevant codebooks in prior IS literature and therefore created my own system.

I applied codes first to transcriptions of fieldwork data in the master Word document. I then, in individual participant spreadsheets, divided codes into two activity types: 1) information-related activities and 2) extra activities (i.e., interactions with participants regarding their research topics or arranging meeting times). Starting with the fieldwork transcriptions provided me with coding experience over my datasets and gave me confidence in later assigning codes to screenshot data, where I focused my analysis on virtual interactions made visible—i.e., the intersection of participants with resources in the GNAE.

4.5.2.1 Interaction Activity Codes: Fieldwork Data

"Extra activities" codes were fairly straightforward and I developed seven codes which encompassed the nature of my interaction with a participant during fieldwork. While these codes do not provide insight into my primary research questions, they provided me with a record of my activities with participants and what we discussed for possible future reference.

One interaction might have activity codes applied to it; for example, discussing when a participant and I might Skype (LOG) and also mentioning a possible participant in the project (PART RECRUIT).

These codes are listed in coded participant fieldnote spreadsheets for each participant with their date of occurrence.

Table 17: Interactive Activity Code Family.

Code	*Definition*	*Occurrence Across Six Participants*
LOG	Logistics: Participant and I discussed next meeting, next steps in research cycle	27
PART RECRUIT	Participation Recruitment: Participant and I discussed how I might find more research participants.	14
REL BUILD	Relationship Building: Participant and I discussed general issues not necessarily related to this project	14
RES TOPIC	Research Topic: Participant told me about their scientific work	13
RES DESIGN	Research Design: Participant and I discussed and/or brainstormed research design concepts	15
PRE-PRINT and PHONE	One-off codes for p1, who talked about how she uses pre-prints generally; I did not find relevant enough to include in the information-activity codes	2 (one of each)

Here is an example of an activity code in action from coded fieldnotes:

> 8. Format: Email
> Participant: p1
> Date: 17 July 2013
> Storage Location: Gmail
> Theme: Names and institutions of recruitment campaign
> Analytic Notes:
> About list: "These are people that I know and have collaborated with. They may still not respond, but perhaps they can forward it to others, which can then lead to some responses—the network effect."
> About email: "I have read over the 'invitation' letter and it is very good.
> Analytic Commentary/CODES:
> **PAR RECRUIT**

4.5.2.2 Information-Related Behavior Codes

Information-related behavior codes center around 1) mentioned or observed information resources and the attributes thereof and—for screenshot/virtual interaction data only—2) activities performed and resource attributes. Information-related codes and attributes relate to RQ1 and RQ2, as will be seen in Chapter Five.

I utilized the aforementioned resource provider master spreadsheets for keeping track of additional information about resource providers I had located online. This kept my central code list manageable and focused on the research questions at hand.

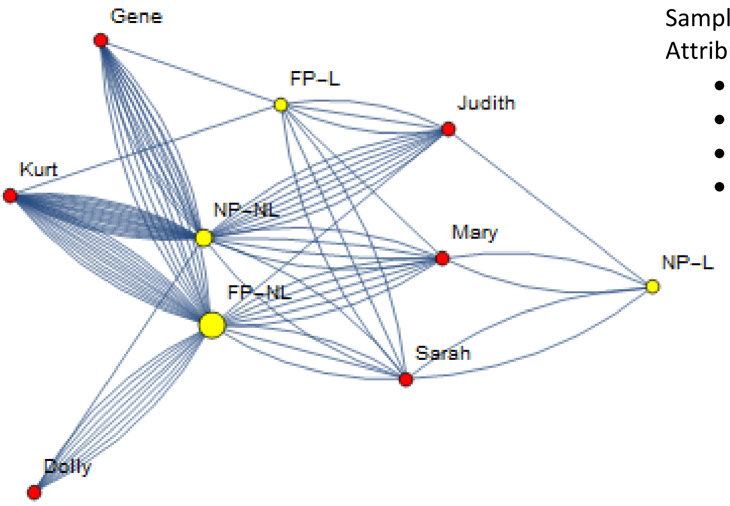
Table 18: Information-Related Behavior Code Family.

Code	*Definition*	*Attributes*
RESOURCE NAME	Variable code; primary name of information resource provider mentioned or observed. Can be abbreviated or full length. Example: GOOGLE US (United States version) COMM NON-LIB	COMM: For-profit provider NON-COMM: Non-profit provider LIB: Resource made available to participant because of identifiable library infrastructure (e.g., subscription, proxy server, repository access). NON-LIB: Library is not involved in providing access to the resource. Note: Optional clarification with journal title or other information in a few cases
ACTIVITY PERFORMED	For screenshot data only; variable code based on activity observed Example: KEYWORD SEARCH "splay rigid"	• Keyword string – ACTUAL STRING ITSELF • Topic or Theme of Item viewed – ITEM TITLE, WHERE RELEVANT

ACTIVITY PERFORMED, Second example	FORM	FORM: Formal activity – formal information type with traditional "authenticated" provider (library, publisher, other) INFORM: Informal activity – informal information type with "unauthenticated" provider *Note: The FORM and INFORM attribute codes turned out to be problematic when I reached the point of interpreting data, as I will discuss in Section 4.5.2.3 directly below, but I left these codes in all coded documents so readers could observe the evolution of codes throughout such a project.*

4.5.2.3 Additional Information Resource Attributes

During the coding process, I discovered I required additional information resource attributes for both fieldwork and screenshot transcriptions because the FORM and INFORM codes above were not satisfactory for capturing who certifies a particular resource as "authoritative" I therefore needed to gain some understanding of this for the interpretation of data in relation to theory. I also wanted to more deeply describe the information resources as "nodes" within the GNAE. For that purpose, I required attributes for resource type (fieldwork and screenshot types identified) and country of origin for resource provider within the GNAE, where available.



Sample Information Resource (node) Attributes:

- *For-profit, Library (FP-L)*
- *Non-profit, Library (NP-L)*
- *For-profit, Non-library (FP, NL)*
- *Non-profit, Non-library (NP, NL)*

Figure 7: Visualization of Concept: Scientists (red nodes) Using Resources (yellow nodes) Within the GNAE (for detail, see Section 5.2.2.1 and Appendix G).

Table 19: Information Resources Attributes Family: Who Certifies Content—Both Fieldwork and Screenshot Transcriptions.

Attribute	*Definition*	*Notes*
Global Review Board	Content is verified in some manner by a global review board	arXiv global review board only
Google	Google/Alphabet Inc. products; information about models of production is highly proprietary	This attribute encompasses all Google products mentioned in fieldwork
Scholarly Publisher	Traditional scholarly publisher organizes and provides information	Elsevier, Thomson Reuters, scholarly journal publishers
Community of Users	Communities of users govern the development and modification of the resource (even if elements of this are proprietary); concept extended to include shareholders for commercial entities	Open-source software, open databases, some commercial resources with collaborative functions
Community of Users + Government	Communities of users are involved, plus an official governmental entity certifies authenticity of the resources	For my data, this was either the United States or European Union government for public databases
Academic Results	For research groups, the relevance of their research results creates trust into the information they provide	I included research groups as information resources for one participant as a way to illustrate the importance of research group information tracking for these participants. I could not include all research groups mentioned because of concerns about identifying very particular sub-specialties

I used resource providers' webpages in combination with Wikipedia US (i.e., publicly-available information about each resource) to create brief definitions. Links to specialized resources are available in the resource provider datasheets.

4.5.2.3.1 Resource Types for Fieldwork

Table 20: Information Resources Attributes Family: Resource Type, Detailed—Fieldwork.

Attribute	*Definition*	*Examples*
LIB: Trad journal	Tradition journal participant has access to via a library subscription	Journal of Cheminformatics
LIB: Science platform	Bibliographic database or other similar platform	Web of Science, Scopus
Citation management	Tool for working with citations in relation to documents	JabRef
Team code management	Tool or platform for teams creating code together	BitBucket
Conference	Conference program or other related material	Cheminformatique Strasbourg
Collaborator	Partner institution for participant	Czech Academy of Sciences
Filesharing	Tool for sharing files in the cloud	Dropbox
LIB: eBooks	General reference; assumes library subscription	(any)
Facebook RSS	RSS provided by publisher directly via Facebook; circumvents library and/or any subscription to the resource	Environmental Health Perspectives on FB alerts
Patents	Patent database or search tool	Google Patents
FDA RSS	One-off code; United States Food and Drug Administration provides RSS directly; no library intermediation	n/a
Code simulating x-ray spectroscopes	Specialized code; one-off descriptive code	FDMNES
Crystallographic programs	Tools to view crystal structures developed for Rietveld analysis of neutrons, nuclear, and magnetic or X-ray powder diffraction	FullProf Suite
Database inorganic crystal structures	Peer-reviewed completely identified structures since 1913; one-off descriptive code	FindIt
Search engine	Traditional commercial search engine; general and not specialized	Google Scholar
Chat	Video or text chat software	Google Talk
Cheminformatics software platform	Visualization and drawing tools for chem- and bioinformatics	ChemAxon
Chemical structure search	Combined chemical and text search engine	chemicalize.org
Data analysis/graphing	Tool for manipulating and understanding data	Open Babel (chemistry-specific)
Research group	Named by participants as exemplary in field	Reymond Research Group

Attribute	*Definition*	*Examples*
Document preparation	Tool solely for preparing academic manuscripts	LaTeX
People identification	Online profile resource	LinkedIn
Programming language	n/a	Python
Big data processing	Distributed computing infrastructure, computing/storage resources	MetaCentrum
Online chemical modeling environment	Chemical database with modeling environment	OChem
Database of chemical molecules and their activities against biological assays	Compound structures and descriptive datasets	PubChem
NON-LIB: Science platform	Bibliographic database or other similar platform that does not require library subscription	PubMed
Open source cheminformatics software	Software development kit for cheminformaticians	RDKit
Discussion forum	Mailing list or archive online	Zinc-fans
Searchable chemistry data, literature and reactions	Chemistry-related, reliable data from various source-types in one place	Reaxys
Nature RSS	RSS provided by publisher directly via Nature website; circumvents library and/or any subscription to the resource (i.e., subscription not required to view content)	n/a
Open source math-sci-engin software	Open source Python software tools	SciPy
Software for modeling liquid surfaces	n/a; one-off code	Surface Evolver
Protein docking software	n/a; one-off code	UDock
Collaborative encyclopedia	Online commonly-edited encyclopedia	Wikipedia; be sure to note language version

To make this list easier to interpret, I then grouped resource type attributes into broader categories.

Table 21: Information Resources Attributes Family: Resources Type, Grouped—Fieldwork.

Attribute	*Definition*	*Examples*
Document preparation	Citation or document preparation tools	Citation management, document preparation
Data analysis	(same as in Table 20)	n/a
Search engine	Extended to include chemical search engine(s)	Google Scholar, chemicalize.org

Attribute	Definition	Examples
Programming language or code	n/a	Programming language, code simulating x-ray spectroscopes
Big data assistance	(same as in Table 20)	MetaCentrum
RSS	Newsfeed, any publisher	FDA RSS
Science platform: library or non-library	Bibliographic database or platform not necessarily provided by library infrastructure	PubMed
Traditional journal	(same as in Table 20)	*Journal of the Physical Society of Japan*
Software, software platform	Resources identified as software which do not fit other categories	SciPy
eBooks and documents	(same as in Table 20)	n/a; participant did not name but referred to generally
Collaborator, research group, conference	Academic partners or resources	Czech Academy of Sciences
Discussion forum	(same as in Table 20)	Zinc-fans
Patent	(same as in Table 20)	Espacenet

Finally, I geocoded resources based on the country of origin; i.e., information provider headquarters or country(-ies) of origin of a resource creator(s)—for example, original developers of open source software or a virtual international team of anonymous contributors.

Table 22: Information Resources Attributes Family: Geocodes—Fieldwork.

Attribute	*Full Country(-ies)*
US	United States
NETH	The Netherlands
CZ	Czech Republic
GER	Germany
FR	France
AUSTRALIA	Australia
ISRAEL	Israel
LUXEMBOURG	Luxembourg
JAPAN	Japan
HUNGARY	Hungary
CH	Switzerland
GER, NETH, AUS, BEL	Germany, Netherlands, Australia, Belgium
NORWAY, CANADA	Norway and Canada
UK, US, JAPAN	United Kingdom, US, Japan
INTL TEAM	n/a
UK	United Kingdom

4.5.2.3.2 Resource Types for Screenshots

I created similar attribute tables for screenshot data.

Table 23: Information Resources Attributes Family: Resources Type, Detailed—Screenshots.

Attribute	*Definition*	*Examples*
LIB: Trad journal	Tradition journal participant has access to via a library subscription	Journal of Cheminformatics
LIB: Science platform	Bibliographic database or other similar platform	Web of Science arXiv
Patents	Patent database or search tool	Google Patent
Search engine	Traditional commercial search engine; general and not specialized	Google Scholar Google US Google CZ
Research group	Named by participants as exemplary in field or illustrated with screenshots	The Kavli Institute of Theoretical Physics
Big Data Processing	Distributed computing infrastructure, computing/storage resources	MetaCentrum
NON-LIB: Science platform	Bibliographic database or other similar platform that does not require library subscription	PubMed PLOS
Open source cheminformatics software	Software development kit for cheminformaticians	RDKit
Discussion forum	Mailing list or archive online	StackOverflow
Collaborative encyclopedia	Online commonly-edited encyclopedia	Wikipedia; be sure to note language version
Blog	Blog	*All Things Metathesis*
Operating System or Manual	Related to operating systems	Linux Documentation Man Page
Product Catalog	Catalog of chemical products	Sigma-Aldrich online product catalog
Government Health Service	Online governmental information	Servizio Sanitario Regionale Emilia-Romagna
Programming Help	Help pages for programming issues	PostgreSQL help

Table 24: Information Resources Attributes Family: Resources Type, Grouped—Screenshots.

Attribute	*Definition*	*Examples*
Traditional library resources or platforms	Journals or platforms provided by a library infrastructure	arXiv, Science Direct
Non-library science resources or platforms	Bibliographic database or platform not necessarily provided by library infrastructure	PNAS
Search Engine	(same as in Table 20)	Google Scholar, Google CZ
Programmer Tools	Operating system or manual, programming help	Linux Documentation Man Page
Patent	(same as in Table 20)	Google Patents
Other	Remaining items, various	Sigma-Aldrich online product catalog

Table 25: Information Resources Attributes Family: Geocodes—Screenshots.

Attribute	*Full Country(-ies)*
US	United States
NETH	The Netherlands
CZ	Czech Republic
CH	Switzerland
UK, US, JAPAN	United Kingdom, US, Japan
INTL TEAM	Unclear
UK	United Kingdom
ITA	Italy
IRAN	Iran

4.5.2.4 Limitations of Coding Process

Even for this relatively small set of data, it is possible I coded some items incorrectly and I discovered and corrected some errors during the coding process. For example, I had originally coded BioMed Central as a non-profit entity, but in reviewing data discovered it was for-profit and updated the datasheets accordingly.

I used publicly-available information provided by resource providers themselves or available on Wikipedia for coding for- or non-profit status as well as countries of origin, so these attributes were current at the time of compiling the data for this dissertation (August 2015) but are subject to change over time (e.g., a company headquarters might move or multiple headquarters be consolidated into one). This, however, should not significantly impact the broader patterns I identify and interpret in Chapter Five.

4.5.3 Information-related Data Analysis, All Participants

4.5.3.1 Summary of Analysis Process

Because I had coded all raw data during the preparation process and placed it in spreadsheet format (in addition to the master Word files), I was easily able to manipulate and analyze coded information-related data, both at the participant level (Section 4.6 below) and at an overall level for both fieldwork and screenshot data. This subsequently allowed me to interpret the data in relation to my central research questions as well as theory.

I will provide here an analysis of data across all research participants for both fieldwork and screenshot data followed by a detailed description of behaviors observed for each participant (see Table 17 above for tallies of the interaction activity code family).

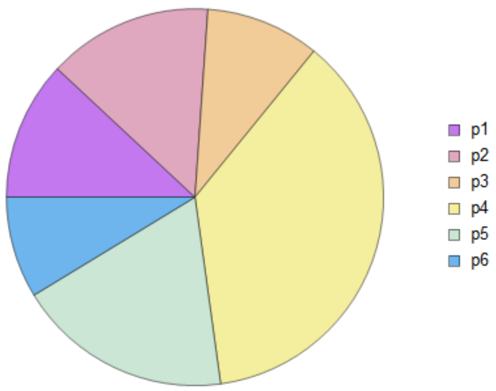
4.5.3.2 Fieldwork Data Analysis, All Participants

I identified seventy-three separate information resources of various kinds mentioned during conventional and virtual fieldwork with the six participants.

I used a broad definition of "information resources" to include not only traditional library subscription databases, but also online research groups, software, programming languages, communication and data analysis tools, discussion forums, specialized databases created by scholars for others in their fields, and even online conference programs—anything which the six participants mentioned. All information resources and tools *are available within the GNAE*, although some require downloading or purchase of program files to local computers (e.g., software programs which do not have cloud versions) for use.

Numbers of resources mentioned during fieldwork varied according to participant. I interacted most with p4, for example, so his list of resources is the longest. Resource lists in this way cannot be considered as representative samples in the quantitative sense (i.e., there is no quantitative "sample imbalance" [Yoon 2004, p. 24]) but rather represent *snapshots of moments in time* spent discussing information exploration and discovery lifeworlds with each participant. Detailed analysis by participant follows in Section 4.6 and I will compare and discuss this data with visual data and in relation to RQ1 and RQ2 in Chapter Five.

Table 26: Number of Resources Mentioned, by Participant.



p1	*p2*	*p3*	*p4*	*p5*	*p6*	***TOTAL***
11	13	9	34	17	8	**92**

No participant mentioned library websites during fieldwork, and no information resource was mentioned by all six participants. This is because I broke Google use into its constituent products and regional versions. Google products/versions mentioned by participants include:

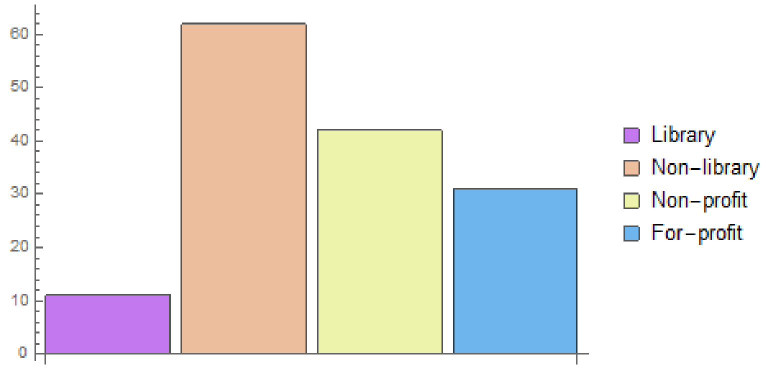
- Google Scholar (five participants)
- Google Patent (two participants)
- Google Czech version (Google CZ, two participant)
- Google (likely Google CZ but not confirmed, one participant)
- Google Drive (one participant)
- Google Hangout (one participant)

In addition to Google, the most commonly-mentioned resources were: arXiv.org (four participants), English Wikipedia (four participants), and Web of Science (formerly Web of Knowledge; three participants). Participants 4 and 5 both mentioned the following resources: BitBucket, Google Patent, RDKit, Skype, and StackOverflow. Participants 1 and 5 both indicated they use Python. All other resources were mentioned only once and—even for the participants working in the same discipline—each participant's list of mentioned resources was unique.

Even though four participants were European, the only non-English resources mentioned during fieldwork were Google's Czech variant and Russian Wikipedia. English Wikipedia was preferred by all participants mentioning it. According to p2, Wikipedia is useful for definitions and "the English scientific articles are much richer than those in [my] native language" (S.Krueger 2013, Fieldwork Transcriptions, p. 7, 29 Jul). All four participants who are not English native speakers publish and present outputs of their research in English and consider English-language resources to be preferable to publications in their own language because they typically have higher impact factors than local journals and publishing in them provides more publishing "credits" for academic promotion.

Across participants, while online library resources are still important—an example being p3, who stated she begins her day "by opening up arXiv in the morning in order to check news and the landscape of what's happening in the field" (S. Krueger 2013, Fieldwork Transcriptions, p. 10, 3 Nov)—they are typically supplemented by a wide variety of online information resources and tools which vary according to research specialties. Of the seventy-three resources mentioned during fieldwork, only eleven were traditional resources provided via a library infrastructure; sixty-two were non-library resources. More resources mentioned during fieldwork are provided by non-profit entities (forty-two) than by commercial organizations.

Table 27: Library Versus Non-library Resources and Tools; Non-Profit, For-Profit.



Library	*Non-Library*
11	62

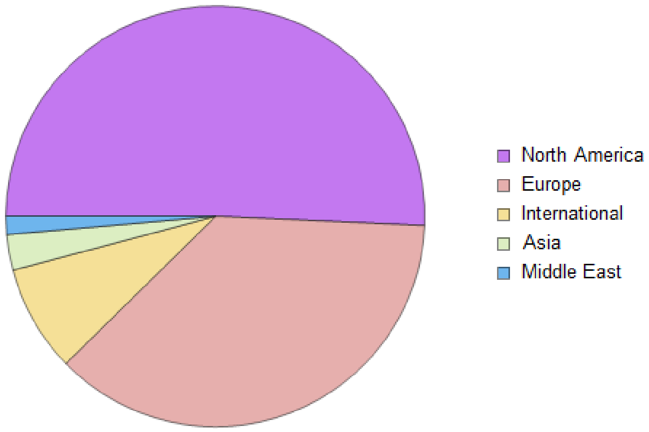
Non-profit	*For-profit*
42	31

Almost half of resources mentioned (thirty-six) are based or were created by organizations in the United States, with twenty-seven resources originating in Europe. Two resources mentioned are provided by institutions in Asia and only one in the Middle East. These numbers include regional collaborators and research groups only for p4 because of privacy concerns I had regarding other participants, whose research is more highly specialized and therefore identities of participants might be traced if I were to reveal their collaboration networks.

Table 28: Information Resource Origin, with Collaborators and Research Groups.

Country or Countries (for multiple)	*Number of Resources*
United States (US)	36
France	5
Netherlands	5 (all Elsevier; different products)
Germany	5
Switzerland	4
Czech Republic	3
UK, US, Japan	3 (all *Nature*; different products)
United Kingdom (UK)	2
Hungary	2 (both ChemAxion; different products)
Australia	1
Israel	1
Japan	1
Canada	1
Luxembourg	1
Germany, Netherlands, Austria, Belgium	1 (European patent headquarters offices)
International Team	1
Norway, Canada	1
TOTAL	**73**

Table 29: Information Resource Origin, by Continent.



Continent (international, for multiple)	*Number*
North America	37
Europe	27
International	6
Asia	2
Middle East	1
TOTAL	**73**

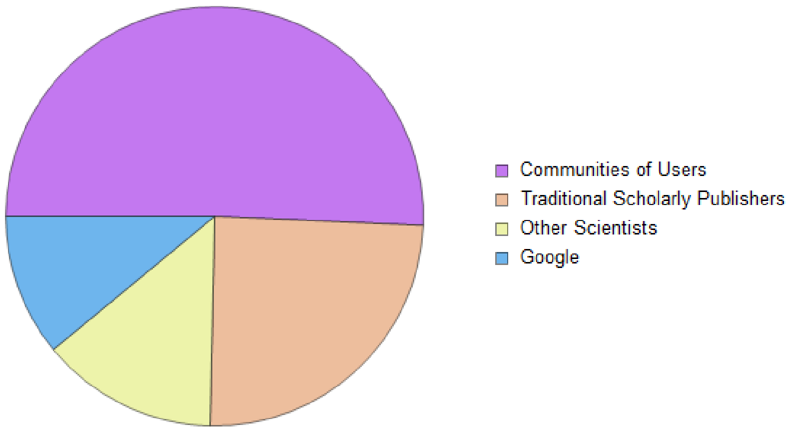
During fieldwork, participants in this study mentioned outputs of other scientists as being useful to them in addition to databases available with or without library intermediation. Additional types of information resources mentioned ranged from data analysis tools to a distributed computing infrastructure p4 and p5 use for processing large amounts of data.

Table 30: Information Resource, by Type.

Type	*Number*
Other scientists (collaborators, research groups, conferences)	13
Database provided by a library *or* beyond library infrastructure	9 (of these, only four are library-provided—either by infrastructure or subscription: arXiv, Web of Science, Scopus, ScienceDirect, and Reaxys)
Traditional scholarly journal	7
Data analysis tool	7
Programming language or code	7
Software or software platform	6
Document preparation	5
Search engine (includes different Google versions plus one chemical search engine)	4
Discussion Forum	4
RSS	4
Chat	3
Collaborative encyclopedia (two language versions of Wikipedia mentioned—English and Ukranian)	2
eBooks and documentation	1
Data processing assistance	1
TOTAL	**73**

Regarding who certifies or authenticates these resources, communities of users (including customers/shareholders for commercial resources, with Google listed separately) are more prevalent than traditional scholarly publishers for the resources mentioned by my participants during fieldwork.

Table 31: Information Resource Authority.



Type	*Number*
Communities of Users	37
Traditional Scholarly Publishers	18
Other scientists (collaborators, research groups, conferences)	10
Google	8
TOTAL	**73**

4.5.3.3 Screenshot Data Analysis, All Participants

Participants provided me with two hundred twenty-nine screenshots (Table 13). Of the forty-two resources illustrated in screenshots, fourteen had also been mentioned during fieldwork.

Of these, only five require library intermediation (WOS, Reaxys, *Nature*, and Science Direct) or are hosted by a library (arXiv).

Eleven exist in the GNAE beyond library paywalls—including the completely open access *Journal of Cheminformatics*.

Figure 8 below presents a 3D visualization of these resources (colored balls/bubbles) grouped according to one aspect of their model of production (non-profit or for-profit) and manner of access (library or non-library). I created all visualizations and graphs in this dissertation using a trial version of Wolfram Mathematica 10.2; for a more detailed description of visualizations and graphs, see Appendix G.

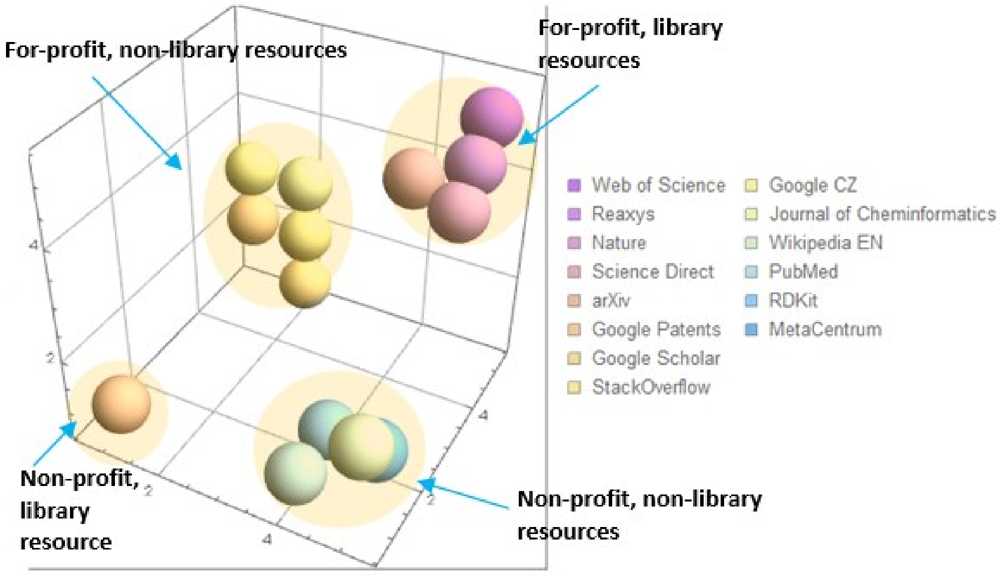
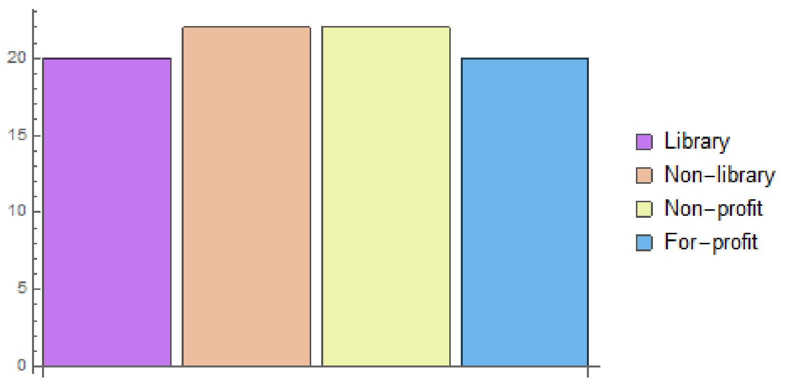


Figure 8: Fourteen Resources Common to Both Datasets.

In screenshots, non-library resources were more common than library resources, but only slightly. As seen in fieldwork, non-profit resources were encountered by participants more than for-profit ones during the exploration and discovery sessions as represented by screenshots. Of the forty-two resource providers identified in screenshot data, twenty-two do not require a library intermediation.

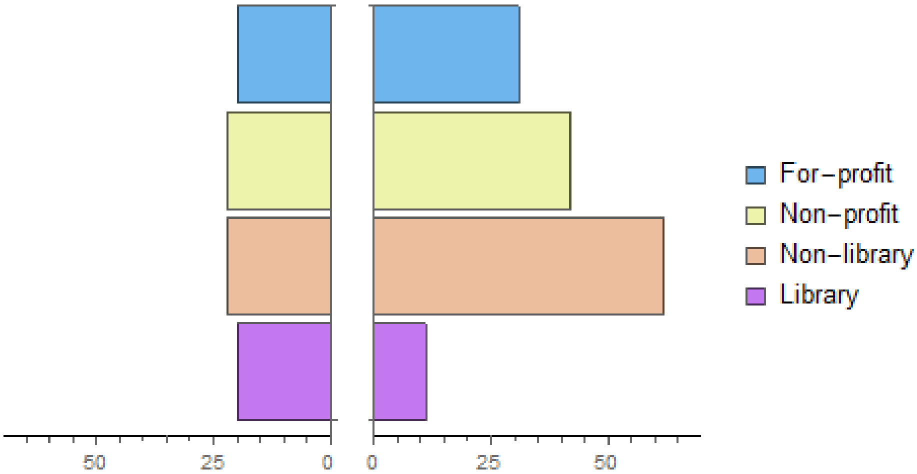
Table 32: Library Versus Non-Library Resources, Screenshots.



Library	*Non-Library*
20	22

Non-Profit	*For-Profit*
22	20

Table 33: Library Versus Non-Library Resources, Tables 30 (fieldwork) and 36 (screenshots) Paired.

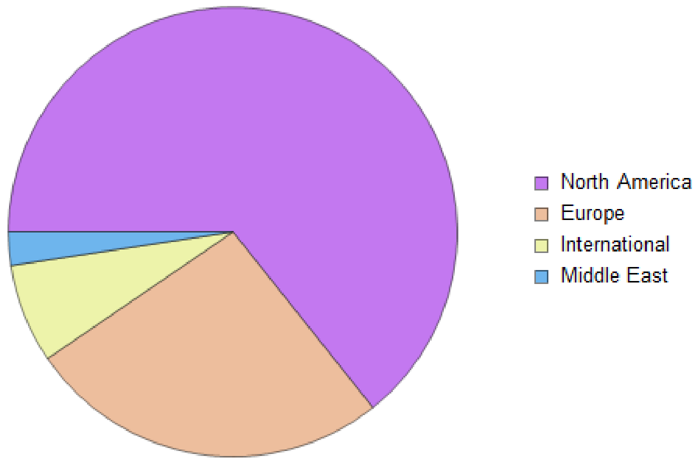


The majority of organizations supporting resources observed in screenshots, as in fieldwork, are based in the United States.

Table 34: Information Resource Origin, Screenshots.

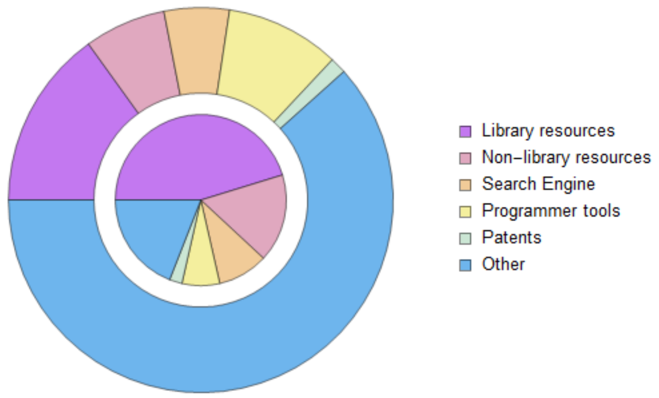
Country or countries (for multiple)	*Number of Resources*
United States (US)	27
UK	4
Netherlands	2 (Elsevier; different products)
International Team	2
Switzerland	2
Czech Republic	2
UK, US, Japan	1 (*Nature*)
Iran	1
Italy	1
TOTAL	**42**

Table 35: Information Resource Origin, by Continent, Screenshots.



Continent (international, for multiple)	*Number*
North America	27
Europe	11
International	3
Middle East	1
TOTAL	**42**

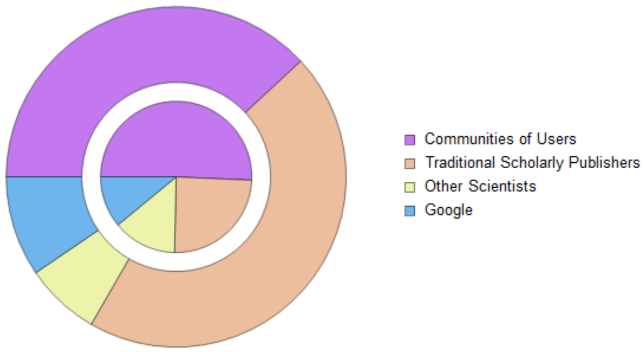
Table 36: Information Resource, by Type: Screenshots (inner pie) v. Fieldwork Mentions.



Type	*Number (Screenshots)*	*Number (Fieldwork)*
Traditional library journals or platforms	20	11
Non-library science resources or platforms	6	5
Search Engine	4	4
Programmer Tools	3	7
Patents	1	1
Other	8	45
TOTAL	**42**	**73**

Twenty resources identified in screenshots are produced via traditional scholarly publishing processes, sixteen are governed by communities of users (including commercial organizations with stakeholders), four by Google, and three by others (e.g., research groups). Numbers for scholarly publishers differ from the "library/non-library" categories in Tables 36 and 37 because of arXiv (governed by a global scientific board though housed in a library) and open access journals which do not require library intermediation for access to them.

Table 37: Information Resource Authority: Screenshots (outer pie) v. Fieldwork.



Type	*Number (screenshots)*	*Number (fieldwork)*
Traditional Scholarly Publishers	19 *(arXiv coded as Communities of Users because of its global review board)*	18 *(higher than "library" resource code because of open access journals which have traditional mechanisms but do not require library intermediation)*
Communities of Users	16	37
Google	4	10
Other (research groups, conferences)	3	8
TOTAL	**42**	**73**

<u>4.5.3.4 Summary of Data Analysis, All Participants</u>

While online information resources intermediated by libraries remain important to participants in this study, a broad spectrum of information-related exploration and discovery beyond library paywalls and traditional scholarly publishing infrastructures is illustrated by fieldwork and screenshot data. According to data across participants in the study, Google in various formats (regional search variants plus Google Scholar and Google Patents) plays an essential role in the information discovery process, with search conducted almost exclusively using English keywords. In both fieldwork and screenshot datasets, more organizations supporting resources are based or were created in the United States than in other countries. While traditional scholarly publishers as "authenticators" of content were important to participants in the study, data from both fieldwork and screenshots indicate that communities of users, Google, and other colleagues are trusted by participants to provide information to them—notable examples of this being arXiv, Wikipedia (English), and Stack Overflow, all governed by virtual global networks of peer communities. Implications of patterns of trust in relation to peer communities will be discussed in Section 5.2.1.2; details about the information-related behaviors of each study participant follow below.

4.5.4 Data by Participant

4.5.4.1 Summary

In this section, I analyze data in relation to each participant, introducing pseudonyms for each participant which I will use for the rest of this dissertation. The roles in Table 43 are the actual roles played by participants at their time of participation in this project.

All six formal research participants had the opportunity to review this dissertation for two weeks in April 2016 prior to publication of this dissertation for 1) accuracy in relation to their research areas and narrative descriptions and 2) to ensure they felt that the level of description did not jeopardize their privacy. Three participants responded with minor comments and updates (p2, p4, and p6), and three participants had no comments (p1, p3, and p5).

Table 38: Participant Pseudonyms.

Participant Number	*Pseudonym*	*Role*
1	Mary Newton	Assistant Professor
2	Judith Ray	PostDoc
3	Sarah Spark	PostDoc
4	Kurt DeSilva	PhD Student
5	Gene Kim	PhD Student
6	Dolly Grant	Associate Professor

4.5.4.2 Mary Newton

4.5.4.2.1 Fieldwork Summary

Prof. Newton is an American condensed matter theorist at mid-career. She has tenure at an RU/H, where she leads a condensed matter theory group. She completed her PhD at an Ivy League institution and received a career advancement fellowship from the US National Science Foundation (NSF) several years ago. She is author of over thirty publications and serves as manuscript referee for five journals and additionally serves as grant proposal reviewer for NSF. Mary is a very committed mentor and is currently supervising three graduate students.

I met Mary through a friend of a friend, and she was an active participant in this project between November 2012 and August 2014.

Mary is committed to producing new research in her field and uses information resources to monitor trends and to support her ongoing research and publishing efforts. She uses LaTeX for preparing article manuscripts and is aware of citation management tools but does not currently use them, although her colleagues have encouraged her to use open source JabRef, a BibTeX reference manager that integrates well with LaTeX.

Mary uses arXiv, Web of Science, and Wikipedia on a daily basis. She collaborates with others primarily using Skype and Dropbox, and occasionally uses her mobile phone during a one-hour commute from her home to her university in order to discuss ideas, including equations, with collaborators.

Mary is aware that her library provides her with access to Web of Science (WOS) and to electronic journals she identifies in WOS but she does not utilize other library services. When asked if she could comment on additional services her library might be able to provide to scholars in her field, she answered only: "I will have to think a little on this one" (S. Krueger 2013, Fieldwork Transcriptions, p.4, 4 Nov).

Regarding library-intermediated arXiv and WOS, she noted that "[m]any condensed matter theorists post their latest work on arXiv, so you can keep up-to-date with what people are working on" and she uses WOS to supplement arXiv: "[it] helps me keep up-to-date with the publications of those who do not submit to arXiv" (S. Krueger 2013, Fieldwork Transcriptions, p.4, 4 Nov).

During fieldwork, Mary mentioned more information resources and tools accessible beyond the library paywall than within it. For example, because Mary is currently extending her collaborative efforts to include biologists in her research efforts, Wikipedia (English version) "is very relevant when I am looking into a new biological systems [sic] that I am trying to model" (S. Krueger 2013, Fieldwork Transcriptions, p.4, 4 Nov).

She uses a variety of commercial and non-commercial tools to assist her in data analysis, graphing, and manipulation, including:

- Mathematica (visualization/computational software program)
- MATLAB (data analysis and graphing)
- Surface Evolver (program for modeling liquid surfaces)
- Xmgrace (2D data analysis and graphing)
- Python (programming language)

The figure below is a visualization of resources named during my fieldwork with Mary, with resources mentioned mapped to a matrix with four categories: commercial library resources, non-profit library resources, commercial non-library tools, and non-profit non-library tools. Of the resources Mary named, five are non-profit, non-library resources and tools. Two are intermediated by libraries (commercial WOS and non-profit arXiv); the rest are commercial, non-library tools and resources.

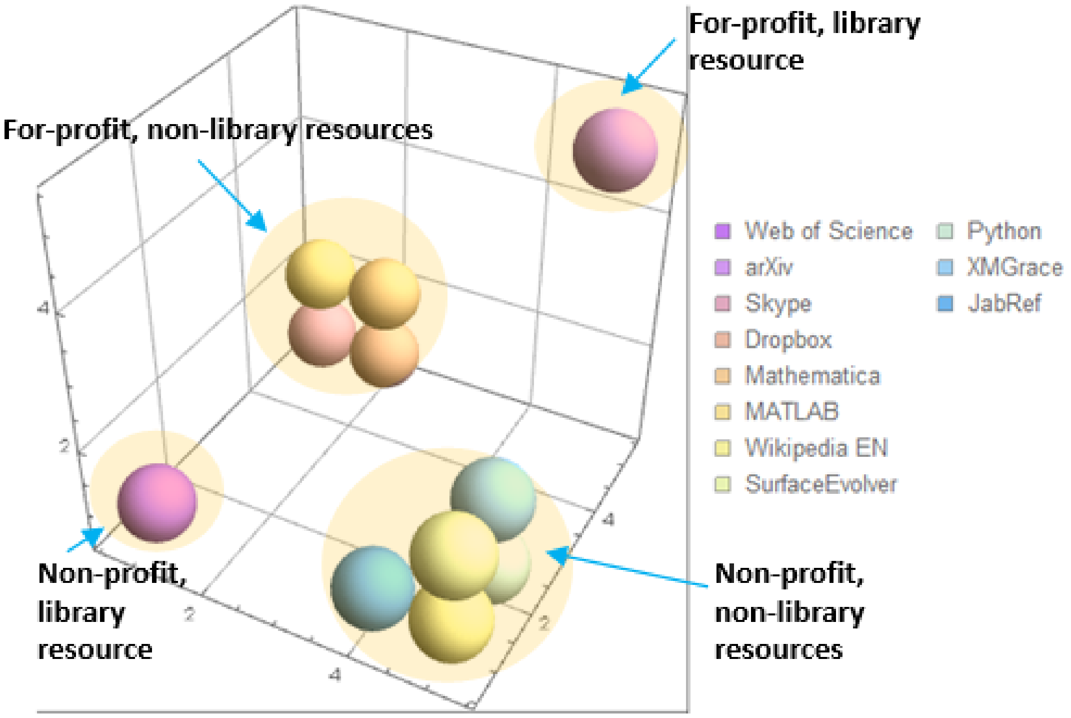


Figure 9: Mary Newton Fieldwork Data Overview.

<u>*4.5.4.2.2 Screenshot Summary*</u>

Mary provided me with thirty screenshots representing her information-related discovery and exploration in the GNAE during two weeks in August 2014. Many of the screenshots represented a traditional pattern of online searching on keywords, browsing search results for relevance, and retrieving article full-text. Mary also successfully conducted author searches in WOS and Scirate.com. No screenshots represented her work with data analysis, graphing, or programming tools. One interaction shows her finding a presentation made freely-available by a research institute.

While Mary did not mention Google as being an important tool during fieldwork discussions, it was her most heavily-used tool (more so than arXiv or WOS) in the screenshot data she provided to me. While her overall patterns of virtual interaction generally match what she reported to me in the field, it was Google—not Google Scholar or a library resource—which typically led her to article full-texts.

One notable interaction represented a failure to reach an article in *Nature*. Although her library subscribes to this journal and although in fieldwork she mentioned she is familiar with her library's remote access mechanisms, Mary was not logged into her library proxy server when she attempted to access the *Nature* article and, upon reaching a paywall, she gave up on getting the full-text and did not proceed to authenticate via her library. She named this screenshot "tried to download" [S. Krueger 2014, Screenshot Transcription, p. 6, 22 August]).

The visualization and table below provide a summary of Mary's information-related behaviors as reflected by screenshot data. In the bubble chart below, the larger circles represent multiple screenshots for a particular resource.

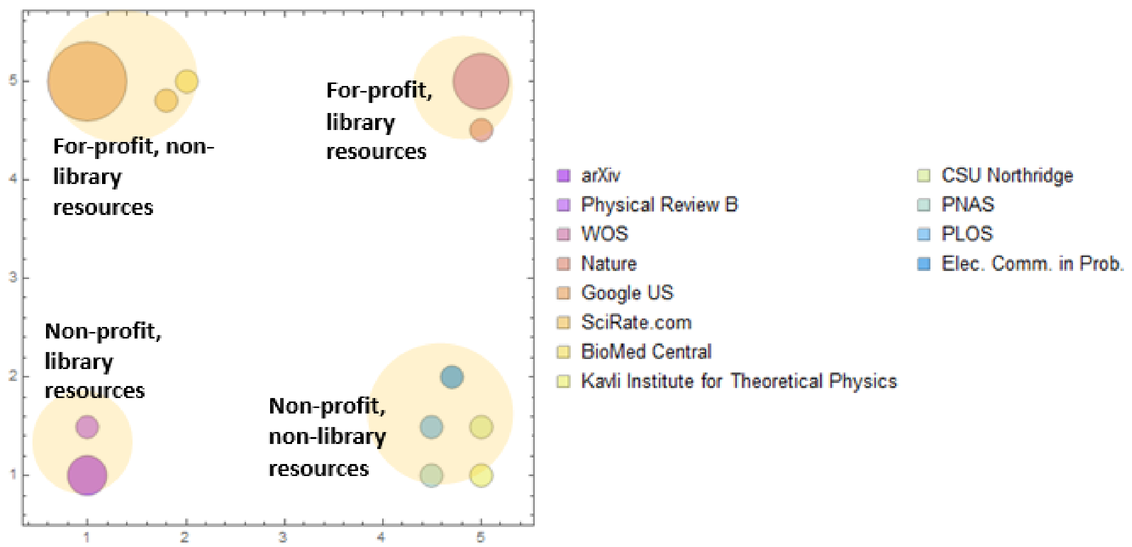
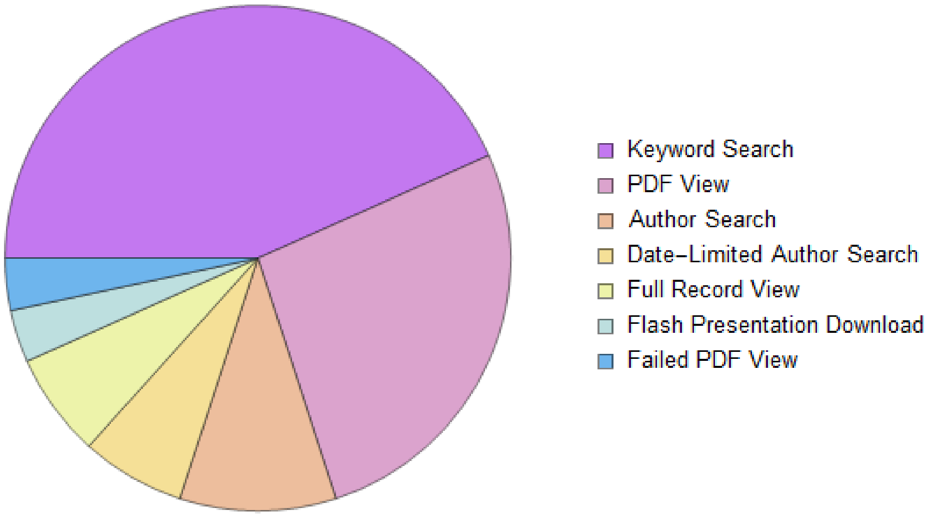


Figure 10: Mary Newton Screenshot Data Overview.

Table 39: Mary Newton Screenshot Activities.



Activity	*Occurrence*	*Source(s)*	*Examples*
Keyword search	13	• Google US • arXiv	"spray-rigid" rigidity percolation continuous transition
PDF View	8	• arXiv • Biomed Central • CSU Northridge (article on a private faculty server • *Electronic Communications in Probability* • *Physical Review B* • PLOS One • PNAS	Central Limit Theorems for the Products of Random Matrices Sampled By a Random Walk Force distribution, multiscaling, and fluctuations in disordered elastic media
Author Search	3	• Scirate.com • WOS	Socolar JE*
Date-limited Author Search	2	• WOS	AUTHOR: (Durian DJ*) AND YEAR PUBLISHED: (1995-1996)
Full Record View	2	• WOS	Green's function measurements of force transmission in 2D granular materials
Flash Presentation Download	1	• The Kavli Institute for Theoretical Physics	Active marginal networks
Failed PDF View	1	• *Nature*	A Levy for a light

4.5.4.3 Judith Ray

4.5.4.3.1 Fieldwork Summary

Prof. Ray was born in the Ukraine and completed her PhD studies at a Russian research university in condensed matter physics, specifically crystallography. Following completion of her doctoral studies, Judith did her first postdoc at a German institution followed by a postdoc at a research university in Prague where I observed her in the field. Judith was conducting experimental research—attempting to grow superconductive crystals—in a laboratory. This research involved uranium, which required her to "ensure that all outputs of material [were] carefully recorded and reported" (S. Krueger 2013, Fieldwork Transcriptions, p. 8, 29 Jul). Our fieldwork ended in late 2013 after Judith's returned to the Ukraine. She was the only participant I did not meet through referral; she and I met randomly in a semester-long Czech class.

Judith's field is highly specialized. Judith mentioned she "uses information resources to both 'get a good impression' of what's going on in [her] field and to find specific information about particular questions [she] has about terminology (e.g., to look up a compound)" (S. Krueger 2013, Fieldwork Transcriptions, p. 7, 29 Jul). Judith also noted she uses databases to "conduct people searches, to gauge the activity of a research group ('usually lots of publications indicate an active group') and to search for publications." She did not know what role the library played in providing access to databases—Judith only knew that she had them, unlike at other times in her career when she did not have access to them and "was forced to ask other researchers who had access to needed journals to download and send articles essential to [her] research" (S. Krueger 2013, Fieldwork Transcriptions, p. 7, 29 Jul).

The last time Judith entered a physical library was 2008, although she said she occasionally uses books because they contain useful crystallographic information that does not become outdated. She prefers printed materials of articles or books to electronic formats because she finds print easier than "flipping through screens; books are not DVDs" (S. Krueger 2013, Fieldwork Transcriptions, p. 7, 29 Jul).

Library subscription resources Judith said she uses include WOS, Scopus, ScienceDirect. An important journal for her was the *Journal of the Physical Society of Japan* (JPSJ) because of its "world class" coverage of uranium-related topics. (S. Krueger 2013, Fieldwork Transcriptions, p. 8, 29 Jul). Judith noted sometimes articles are easier to find using Google Scholar (GS) than subscription databases and cited an example where she found an article using GS for “looking up a compound with a non-English search term” which would not have been indexed in WOS or Scopus (S. Krueger 2013, Fieldwork Transcriptions, p. 8, 29 Jul).

Like Mary Newton, Judith mentioned she uses arXiv to “see what’s happening” and Wikipedia for definitions—she noted “the English scientific articles are much richer than those in [her] native language” (S. Krueger 2013, Fieldwork Transcriptions, p. 8, 29 Jul). She, as Mary, writes articles in LaTeX and analyses data using a commercial tool called OriginLab.

Also important for Judith were three non-commercial, non-library specialized tools: FDMNES, code from the French NEEL Institute which simulates x-ray spectroscopies linked to real absorption or resonant scattering of synchrotron radiation; FindIt, a peer-reviewed database of completely identified inorganic crystal structures which have been discovered since 1913 created by the Leibniz Institute for Information Infrastructure in Karlsruhe, Germany; and FullProf Suite, a set of specialized crystallographic tools for working with x-ray and neutron data.

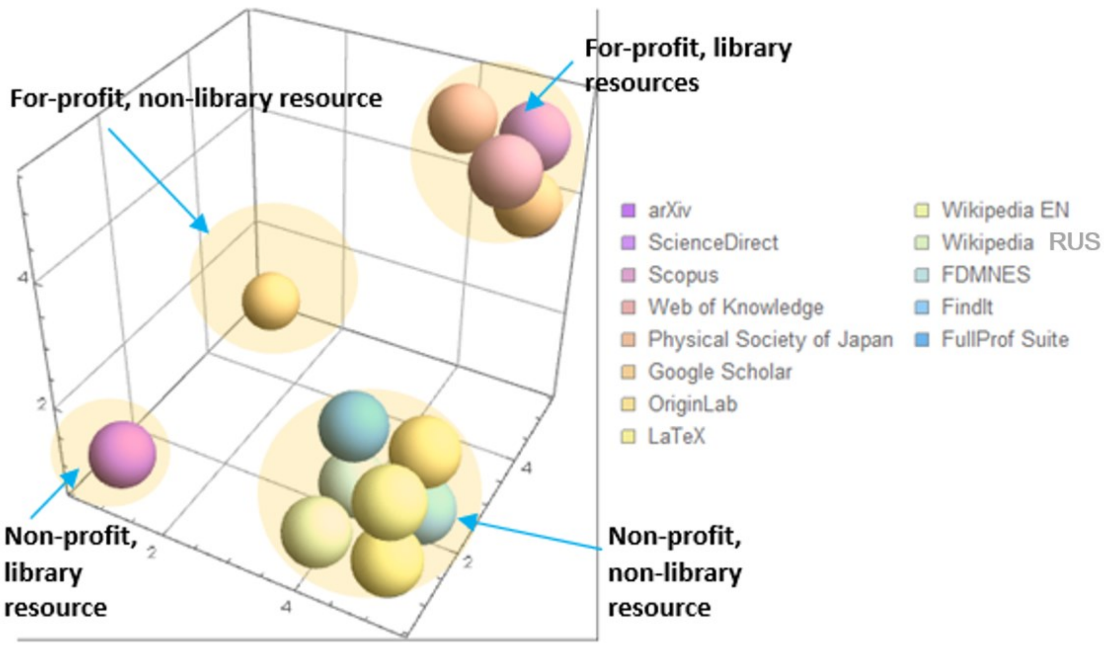


Figure 11: Judith Ray Fieldwork Data Overview

4.5.4.3.2 Screenshot Summary

At the end of our research cycle, immediately before her return to the Ukraine, Judith provided me with three screenshots of her interactions with the GNAE. These screenshots illustrate traditional online resource searching and browsing behaviors in resources made available by libraries—her own university library, for WOS and Science Direct; Cornell University Library for arXiv.

One screenshot showed Judith browsing arXiv for condensed matter journals, and two were keyword searches: one in ScienceDirect for UCoGe, which is a ferromagnetic superconductor, and one in WOS for CeCOIn5, a heavy-fermion superconductor with a layered crystal structure (S. Krueger 2013, Screenshot Transcriptions, p. 9, 11 Sept).

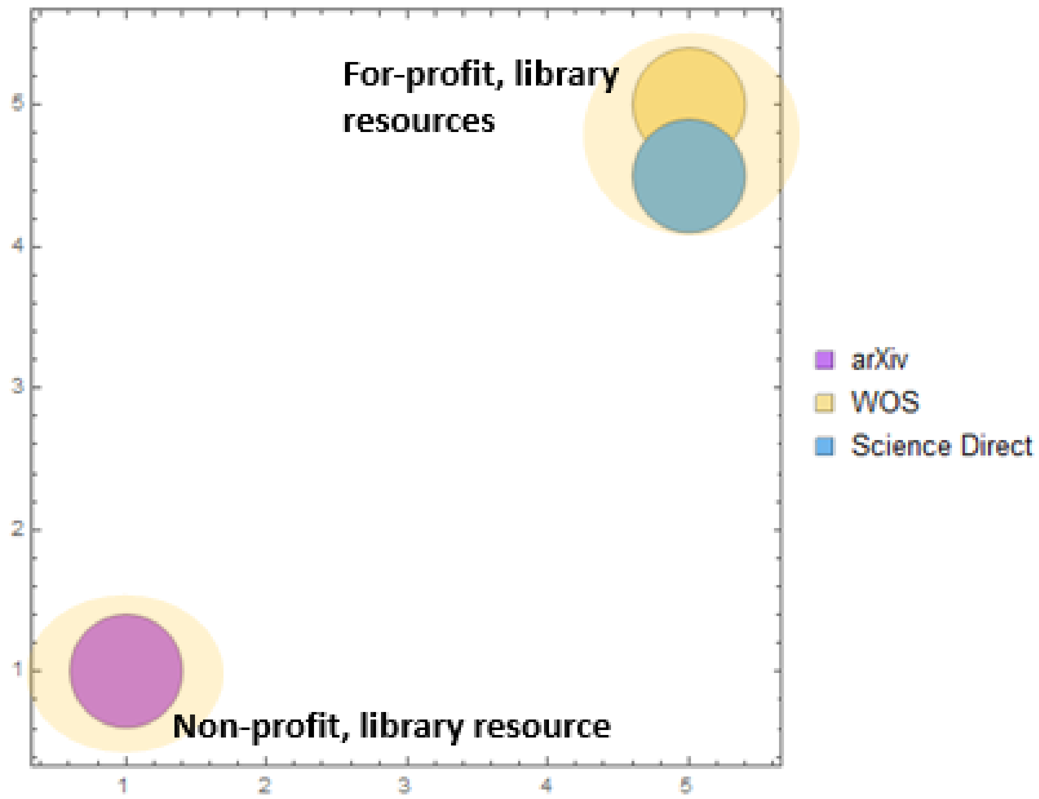
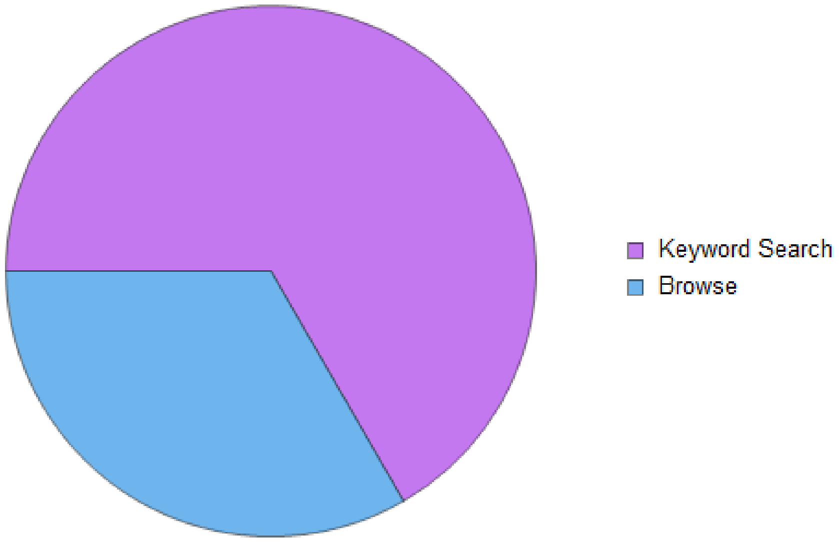


Figure 12: Judith Ray Screenshot Data Overview.

Table 40: Judith Ray Screenshot Activities.



Activity	*Occurrence*	*Source(s)*	*Examples*
Keyword search	2	• Science Direct • WOS	UCoGe CeColn5
Browse	1	• arXiv	Condensed Matter New submissions

4.5.4.4 Sarah Spark

4.5.4.4.1 Fieldwork Data

Prof. Spark was conducting postdoc research in experimental condensed matter physics in the same research group with Judith when I conducted fieldwork with her. Sarah was born in the Czech Republic and is now professor at the laboratory where she did her postdoc. She was the only participant who did not provide me with screenshot data, so I discuss here only the resources she mentioned during a brief period of conventional fieldwork (one hour in-person plus three virtual interactions).

Sarah scanned trends and found articles in a manner similar to Mary and Judith. She noted she starts every day reviewing arXiv. WOS is “an important source of finding articles and tracking citations” although in WOS, “things might be several months behind” (S. Krueger 2013, Fieldwork Transcriptions, p. 10, 3 Nov). She said prefers Google to Google Scholar because she finds the latter gives her “too many irrelevant results” although one might expect this to be the opposite, and she has a set of specific resources and journals she reads monthly, including *Science*, *Nature*, the Physics APS portal, and APS's *Review of Modern Physics*, which has—in Sarah's opinion—"the most impact factor" (S. Krueger 2013, Fieldwork Transcriptions, p. 10, 3 Nov).

Regarding traditional library services, she finds her institution's library catalog to be inadequate and "sometimes ends up purchasing books [herself or through her research group leader]" (S. Krueger 2013, Fieldwork Transcriptions, p. 10, 3 Nov). As in Judith's case, Sarah mentioned older books are helpful in her field "because they are well-written and the basics don't change" (S. Krueger 2013, Fieldwork Transcriptions, p. 10, 3 Nov). Sarah found it difficult for her to imagine how libraries might assist her beyond providing access to materials because "the subject areas are so specific"—in her research group, senior researchers and not librarians "show students how to use research tools, how to read articles, and how to find them" (S. Krueger 2013, Fieldwork Transcriptions, p. 10, 3 Nov).

As Judith and Mary, Sarah writes articles in LaTeX. She uses Reference Manager/BibTex for citation management and occasionally collaborates with others using Skype and email.

Sarah did not mention data analysis or specialized databases in her discussions with me.

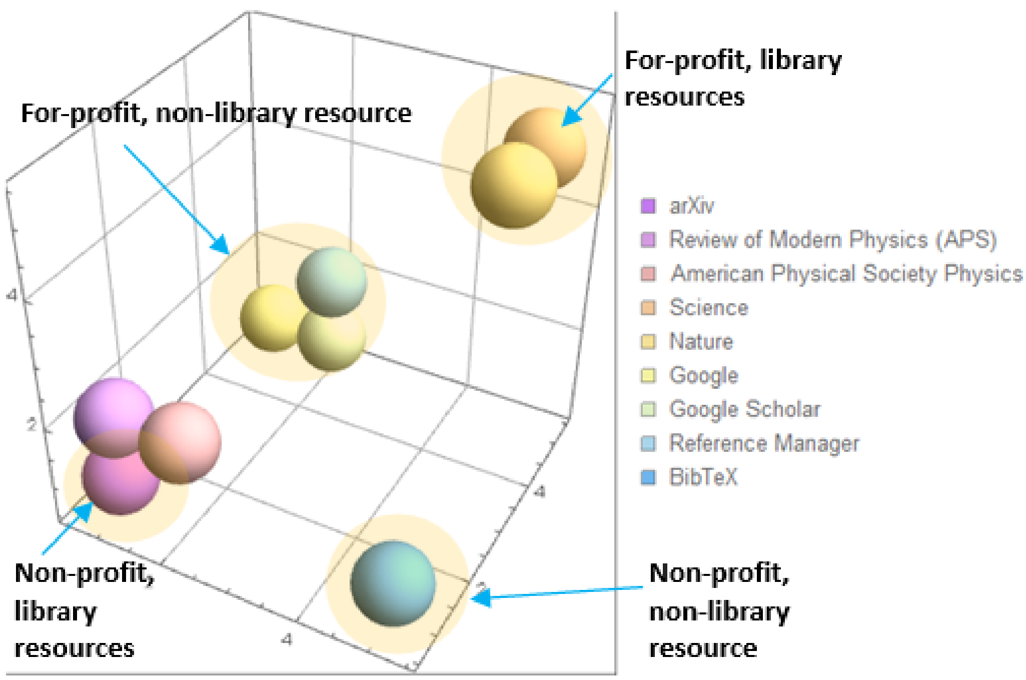


Figure 13: Sarah Spark Fieldwork Data Overview

4.5.4.5 Kurt DeSilva

4.5.4.5.1 Fieldwork Data

MSc. DeSilva, of Czech origin, is currently a doctoral student in bio/cheminformatics at a research university. He is working part-time on a collaborative research project with the Czech Academy of Sciences. Bio- and cheminformatics are:

> Relatively new fields that specialize in application of computer science and techniques on the problematics of life sciences. Currently, [Kurt is] trying to test the viability of *in-silico* chemical space exploration for virtual screening of potential bioactive compounds. (S. Krueger 2014, Fieldnotes Transcription, p. 11, 29 May)

Kurt considers himself to be a theorist, noting:

> The theorist vs experimentalist distinction is a bit difficult with cheminformatics: from the viewpoint of informatics alone, we are experimentalists, because we code tools and process actual data. From the viewpoint of natural sciences (the whole chemistry part), we are firmly in theorist territory. In context of your study (research interests of other participants), we are certainly **theorists**, because our work does not directly involve any tangible processes or products. We just shuffle electrons in processors instead of test tubes :) (S. Krueger 2015, Pers. Corres. 28 July).

Of all participants in this study, I had the most interaction—including virtual—with Kurt, and consider him be a key informant for this project.

Because Kurt's research spans disciplines, he "can't fixate on just a few specific sources" and he needs to keep aware of trends and does so by utilizing RSS feeds "to passively keep up with general important stuff," including *Nature's* general news and drug discovery feeds as well as the open source *Journal of Cheminformatics* (S. Krueger, 2014, Fieldwork Transcriptions, p. 11, 29 May). He did mention using WOS and PubMed, but Google's various products are important to Kurt because they are "free and convenient" and offer "links to other, more specialized sources [through a] clear, unified interface, so [he doesn't] immediately deal with paywalls, forms, registration offers, etc. GS is the usual first stop when I look for relevant papers" (S. Krueger, 2014, Fieldwork Transcriptions, p. 11, 29 May). He noted he never needs print books or other library services, but he is aware the library "pays [for] and manages subscriptions for commercial databases and journals, both of which [are] very useful for me as sources of data and methodology" (S. Krueger, 2014, Fieldwork Transcriptions, p. 11, 29 May). Instead of print books, he uses electronic books and other electronic publications, because they "generally and systematically cover large topics like 'molecular descriptors' or APIs of a tool. Unlike paper books, they are more available, can be string-searched and kept on alt-tab" (S. Krueger, 2014, Fieldwork Transcriptions, p. 11, 29 May).

Kurt is keenly aware of the open access movement. The top journal in his primary field, the aforementioned *Journal of Cheminformatics*, is completely open access. Because his research involves programming, open source tools, forums, and databases are highly-valued and respected—and used on a daily basis. What matters is that open sources provide "practical solutions" to practical problems involved "writing code and processing data"—with "specialized programming libraries" being "essential" to Kurt's work, as are "repositories with version control" which "are the industry standard for coordinated project effort" (S. Krueger, 2014, Fieldwork Transcriptions, p. 11, 29 May).

Wikipedia (English version), as with other participants in the study, is valuable to Kurt. It is:

> Very useful to get a foot[hold] into problematics that [he has] absolutely no clue about. Quality of individual articles [vary], but wiki usually yields basic, easily understandable info with links to more scientific sources. (S. Krueger, 2014, Fieldwork Transcriptions, p. 11, 29 May)

Publisher-provided RSS feeds/alerts are also useful to Kurt, who noted he uses them to "passively keep up with general important stuff" on "general, scientific, and bio/cheminformatic topics" because they have "current and relevant informations [sic] with the best effort/info ratio" (S. Krueger, 2014, Fieldwork Transcriptions, p. 11, 29 May). He also uses feeds to monitor interesting projects.

As was the case with the three physicists above, Kurt writes papers in LaTeX. He uses EndNote, provided by his institution, for managing citations, which he exports to BibTeX; this is "quick and not as error-prone as typing the references manually" (S. Krueger, 2014, Fieldwork Transcriptions, p. 11, 29 May). He would find it helpful if a library would help him in "communicating with editors and getting article/style guides and standards for publishing" (S. Krueger, 2015, Fieldwork Transcriptions, p. 16, 11 Mar).

For collaboration, Kurt uses Google Talk "for its convenience. Faster, more interactive than email while not as obtrustive as skype. However, [he prefers] email for serious communication and requests" (S. Krueger, 2014, Fieldwork Transcriptions, p. 12, 29 May).

RDKit, an open source cheminformatics software and "the best cheminformatics library in my completely unbiased opinion ;)" is essential for Kurt's research, as is the Czech

MetaCentrum, a national distributed academic computing infrastructure upon which his research group is running their cheminformatics project, a CPU-hungry software for exploring chemical space (S. Krueger, 2014, Fieldwork Transcriptions, p. 13, 17 Jun).

Online conference materials and outputs of other research groups are important to Kurt for following trends and for identifying potential contacts for questions and collaboration. Research groups include both those at other universities as well as commercial firms, such as Novartis, because they “do a lot of cheminformatics in the commercial sphere” (S. Krueger, 2014, Fieldwork Transcriptions, p. 13, 17 Jun).

When looking up patents, Kurt prefers Google Patents to other tools and noted “Google does it better” particularly in finding patent owners, since it “[back-refs] to legal events” (S. Krueger, 2014, Fieldwork Transcriptions, p. 15, 21 Aug).

As with other participants, data analysis is a necessary part of Kurt’s work. For this purpose, he uses open source KNIME.

The following figure provides an overview of Kurt’s information resource use as represented in fieldwork conversations. In this figure, I removed research groups for better-comparison with other participants. The majority of tools he mentioned do not require library intermediation. Only WOS requires a library subscription for access.

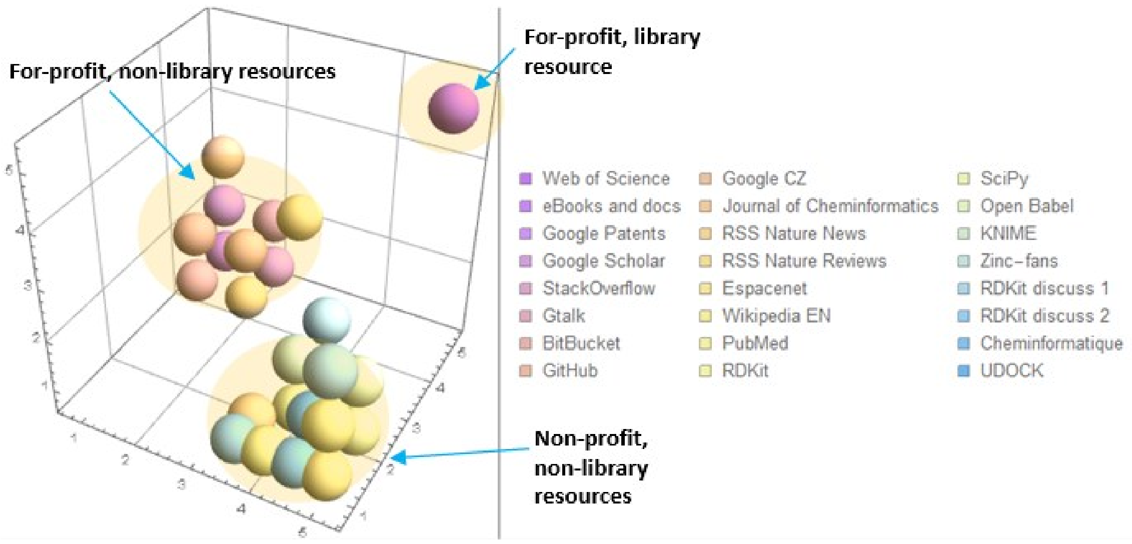


Figure 14: Kurt DeSilva Fieldwork Data Overview (research groups excluded)

<u>*4.5.4.5.2 Screenshot Data*</u>

Kurt provided me with the most screenshots—one hundred and twelve—of any participant in the study, with the majority representing keyword searches in Google (Czech, non-Scholar version) on specific topics. Kurt's data was the most varied of any participant and included a commercial chemical substance product catalog (Sigma-Aldrich), a blog (*All Things Metastasis*), online documentation (PostgreSQL), Wikipedia (English), StackOverflow, RDKit, the Journal of Cheminformatics, PubMed, ChEMBL, and finally vitual interactions with MetaCentrum.

Kurt also provided me with a description of activity for each screenshots, which provide rich detail about what he was doing for each virtual interaction (Appendix G). I asked all other participants for such contextual information, but was unfortunately unsuccessful in obtaining it from others.

Kurt interacted with two only library-provided resources during the time he created screenshots: Reaxys and the *Journal of Investigative Dermatology*, a *Nature* group journal to which his university's library subscribes.

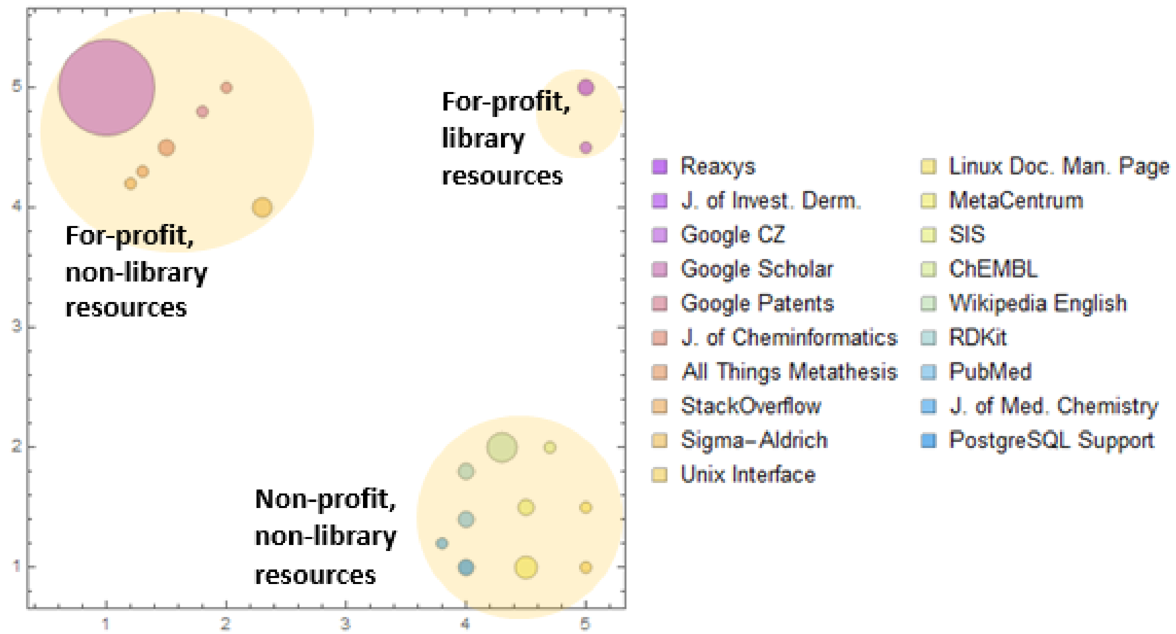
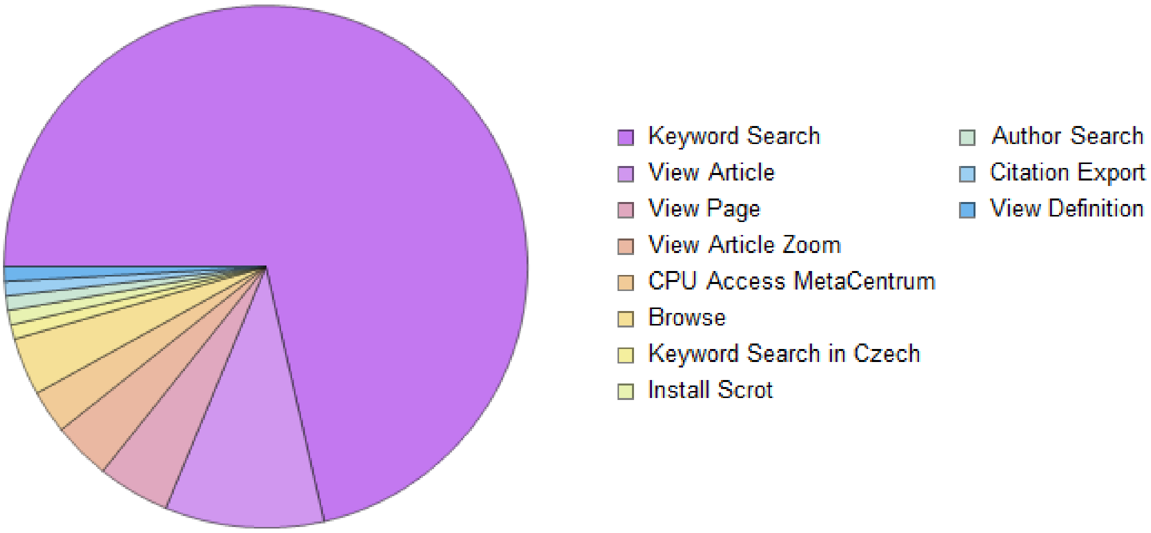


Figure 15: Kurt DeSilva Screenshot Data Overview

Table 41: Kurt DeSilva Screenshot Activities.



Activity	*Occurrence*	*Source(s)*	*Examples*
Keyword search	80	• Google CZ	methyl triphenyl phosphonium bromide python postgresql
View Article	11	• *Journal of Investigative Dermatology* • Wikipedia (English) • *Journal of Medicinal Chemistry* • PubMed • Sigma-Aldrich • *All Things Metathesis*	Rational ligan-based virtual screening and structure-activity relationship studies in the ligand-binding domain of the glucocorticoid receptor-α. Methyltriphenylphosphonium bromide Metathesis: Ruthenium-Based Metathesis Catalysts
View Page	5	• StackOverflow • PostgreSQL Database Adapter Support	Metathesis: Ruthenium-Based Metathesis Catalysts Select rows which are not present in other tables Bug #412000; KDS comment: *The time-honored tradition of copying an encountered programming error into google, verbatim ☺ It usually works.*
View Article Zoom	4	• Wikipedia (English)	Dexamethasone GR knockout mice; KdS comment: *Looking for what happens to mammals without working glucocorticoid receptor.*

CPU Access MetaCentrum	3	• MetaCentrum	n/a
Browse	4	• PostgreSQL documentation • Student Information System • *Journal of Cheminformatics*	*Journal of Cheminformatics – Latest Articles*
Keyword Search in Czech	1	• Google CZ	spravna vyrobi praxe (*good manufacturing process*)
Install Scrot	1	• Unix interface	KdS comment: *After installation, I tried to run the command (the last role of the console) and it snapped this screenshot of the whole desktop.*
Author Search	1	• Reaxys	"sindelar, vladimir"
Citation Export	1	• Reaxys	134 citations
View Definition	1	• *Journal of Cheminformatics*	definition of chemical similarity

4.5.4.6 Gene Kim

4.5.4.6.1 Fieldwork Data

MSc. Kim shares an office with Kurt and is working on the same research project. He is also a doctoral student and is one year ahead of Kurt in this endeavor. Gene, who is Czech, describes his field as "cheminformatics, chemical space mapping, and analysis of chemical databases" and, like Kurt, he is aware of open solutions (S. Krueger 2014, Fieldwork Transcriptions, pp.22, 9 Jul). While the library "is necessary for accessing articles as there are not many open access journals" and while "[i]n our field, chemical databases are also important and, of course, mainly commercial," Gene noted he tries to use open solutions whenever possible (S. Krueger 2014, Fieldwork Transcriptions, pp.22-23, 9 Jul). Regarding books, Gene noted "[s]ometimes, it is useful to have a (e-)book which covers a larger area of research and is more informatics than review papers" (S. Krueger 2014, Fieldwork Transcriptions, p. 23, 9 Jul).

Gene noted he uses LibreOffice instead of LaTeX, and is not currently using citation software.

Google Scholar is a key resource for Gene—it has "the most convenient and fastest searching, for documentation of software, programming languages and even for journal articles (S. Krueger 2014, Fieldwork Transcriptions, p. 23, 9 Jul). He said he uses Wikipedia

(English) “for general information, introduction to a topic” and the StackOverflow programming forum, because his “questions are not unique and almost all questions [are] asked and answered” (S. Krueger 2014, Fieldwork Transcriptions, p.23, 9 Jul).

Library-provided resources Gene mentioned using include Web of Knowledge (WOS), for “special journal queries, # of citations, # of author’s articles, etc.” and Reaxys, for “checking compounds and reactions” (S. Krueger 2014, Fieldwork Transcriptions, p. 23, 9 Jul). Gene noted he complements Reaxys with non-library chemical databases, including Chemicalize, OChem, and PubChem.

RDKit is, as for Kurt, crucial to Gene’s research. For version control, he stated he uses BitBucket and Eclipse as an IDE (Intergrated Development Environment). For drawing molecules, ChemAxon’s MarvinSketch is helpful as is Standardizer, which helps transform chemical structures into representations for use in chemical databases (S. Krueger 2014, Fieldwork Transcriptions, p.23, 9 Jul).

As in Kurt’s case, Gene mentioned he prefers Google Patents to other patent searching tools and noted its summarizing tool (“US, world, EU”) is particularly useful (S. Krueger 2014, Fieldwork Transcriptions, p.25, 21 Aug).

Gene also said he uses arXiv, as was the case with Judith, Mary, and Sarah. As in the case of Kurt, Gene reported use of more non-library resources and tools than those intermediated by a library, with eight of the non-library resources being commercial product and seven being non-profit.

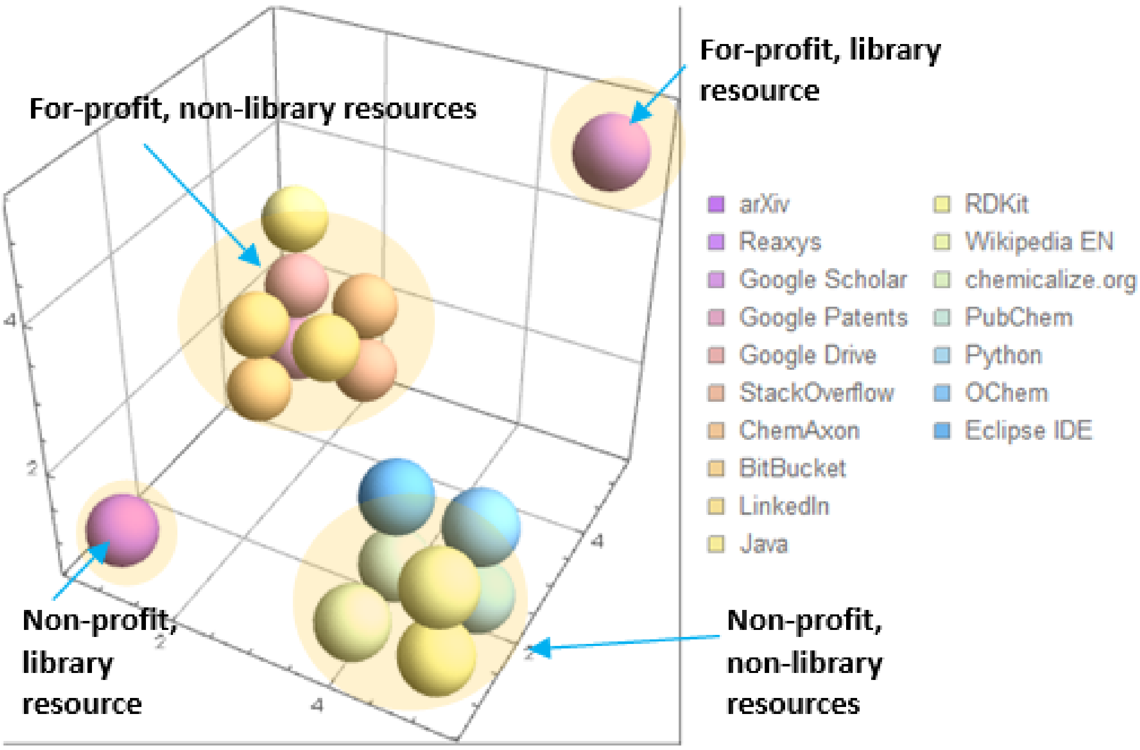


Figure 16: Gene Kim Fieldwork Data Overview

4.5.4.6.2 Screenshot Data

Gene provided me with eighteen screenshots illustrating use of two resources: Google (Czech non-Scholar version) and Reaxys. The Reaxys screenshots are interesting from the point of virtual interaction, because they illustrate structural searches, in which the user enters a visual structure sketch which is then matched across the Reaxys database.

As in Kurt's case, many of the keyword searches relate to specific programming questions.

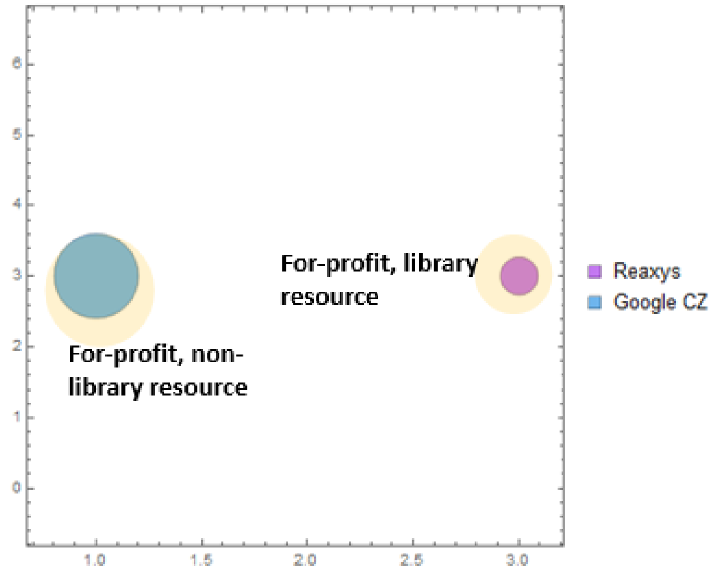
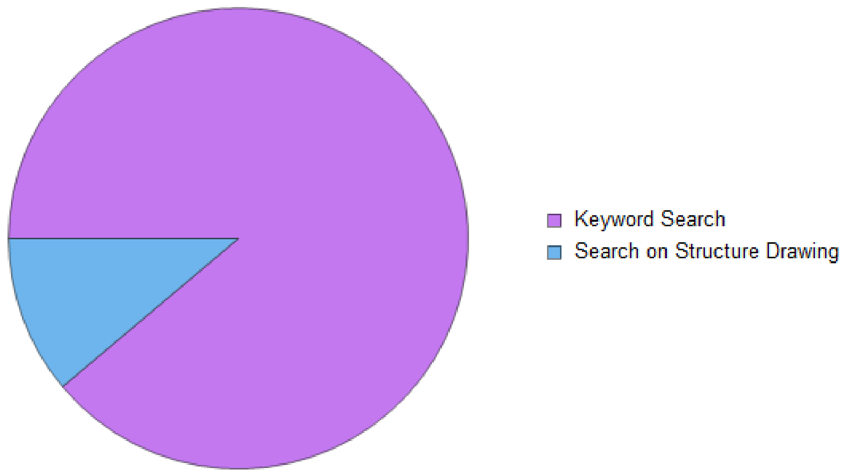
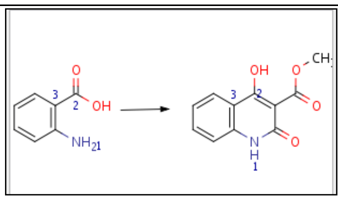
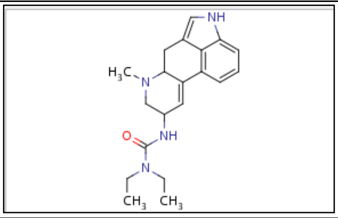


Figure 17: Gene Kim Screenshot Data Overview.

Table 42: Gene Kim Screenshot Activities.



Activity	*Occurrence*	*Source(s)*	*Examples*
Keyword search	16	Google CZ	sql select random rows brufen wiki
Search on Structure Drawing	2	Reaxys	

Activity	*Occurrence*	*Source(s)*	*Examples*
Search on Structure Drawing			

<u>4.5.4.7 Dolly Grant</u>

<u>*4.5.4.7.1 Fieldwork Data*</u>

Prof. Grant is an American Associate Professor at an RU/VH. She, like Mary, is at mid-career and has an extensive publishing history dating back to the 1990s. She is an experimentalist in toxicology (environmental and clinical) in a department of immunology and infectious diseases. I conducted all fieldwork with Dolly virtually, and she is the most recent participant in the project, becoming a formal participant in April 2015.

Dolly, like Kurt, noted she uses publisher alerts to stay abreast of trends in her area of research. She was the only participant in this study who noted she uses alerts publishers place on Facebook (FB). She stated: "I like FB alerts when I have down time, I can see exciting research articles on EHP [*Environmental Health Perspectives*] or the *epigenetics* sites" (S. Krueger 2015, Fieldwork Transcriptions, p. 27, 19 May). She also uses the US Federal Drug Administration's (FDA) alert service, because it:

> Provides daily update on opportunities with FDA. I get a 'smarts' daily grant alerts [sic via email. This is open to all key words. Problem, is that this takes time to sort through. Advantage is that I learn funding patterns across disciplines. When I have time, I value learning about funding patterns and unique opportunities I can apply my research skills to, that otherwise would not be available by a restrictive key word search. (S. Krueger 2015, Fieldwork Transcriptions, p. 27, 19 May)

Dolly noted she begins a search for information with Google Scholar and uses the library's "VPN to search our library resources when I perform a google scholar search" and she is aware of open access resources as well (S. Krueger 2015, Fieldwork Transcriptions, p. 27, 19 May). She believes libraries can assist scholars in becoming visible and stated "[w]ith the development of an institutional repository, we [faculty members] now have the opportunity for our research to become more available to a wider audience, used, and cited" (S. Krueger 2015, Fieldwork Transcriptions, p. 27, 19 May).

Dolly said she uses SAS JMP software for data analysis, and Skype, Dropbox, and Google Hangout ("but not much") for collaborating with research colleagues.

Dolly did not specifically mention any library resources as being useful to her during virtual fieldwork. Although her library subscribes to *Environmental Health Perspectives* and *epigenetics,* she does not use a library-mediated feeds to access alerts but, as noted above, goes directly to FB for this information. These feeds are included as non-library sources in the visualization below.

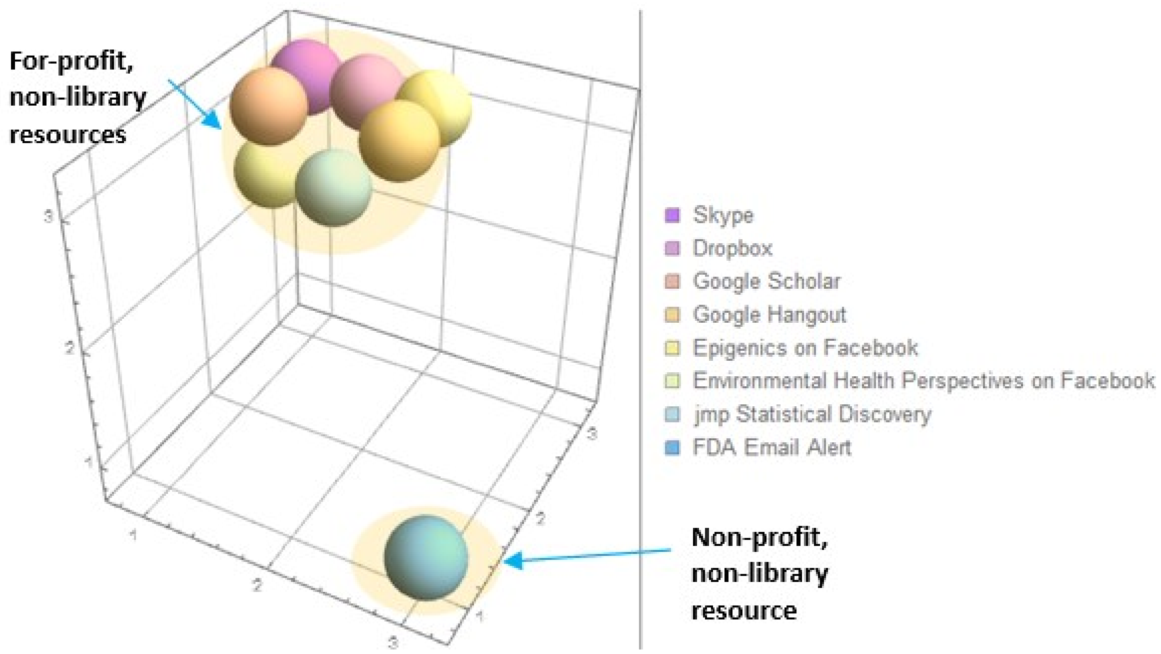


Figure 18: Dolly Grant Fieldwork Data Overview

<u>*4.5.4.7.2 Screenshot Data*</u>

Dolly's sixty-six screenshots represent, in her words, preparing for "[p]ublication—this all day search generated 2 paragraphs for a manuscript that will describe silica and health effects" (p6 2015, pers. comm., 10 Aug). Her screenshot behaviors correspond to what I observed in virtual fieldwork, with Google Scholar being a starting point for keyword searches followed by reading abstracts for relevancy and attempts to reach full-text articles.

As in Mary Netwon's case, although Dolly mentioned to me she knew how to access library resources remotely during fieldwork, she was actually not connected to a library proxy or VPN during her screenshot-taking session. She also did not have Google Scholar's library links to her institution enabled. This meant that Dolly, according to screenshot data, encountered publisher paywalls for resources to which her library subscribes and she therefore extra time attempting to find article full-texts, often in vain. If she had been logged into her library during the searching session and/or had correct library settings in Google Scholar, Dolly would not have had as many difficulties as she did in reaching article full-texts.

Dolly utilized PubMed Central quite extensively to complement keyword search results in Google Scholar, and she also utilized Europe PubMed Central, and had several article successful full-text article views beyond the library paywall for items which *are openly accessible.* For example, she viewed an older open access article in the *American Journal of Respiratory Cell and Molecular Biology,* a journal which requires a library subscription/authentication for access to current issues full-text. Dolly had one incidence of viewing data accompanying an article (she was "learning about" the data, p6 2015, pers. comm., 10 Aug) and attempted to gain full-text at Science Direct, Maney Publishing, Biomed Central, Taylor and Francis, Sage, and directly through journal publishers.

Dolly's screenshot data illustrate how confusing the GNAE can be for scientists when encountering paywalls when they are working outside IP address ranges of their libraries as well as the importance of open access resources, including those funded by national or cross-national governments such as PubMed Central (which I have included in the non-library category since it is not provided by the institution at which she works but rather by the US National Institutes of Health) or Europe PubMed Central, for providing unrestricted full-text to scholars in health-related fields. I have indicated which resources with which Dolly interacts are open access ones in Table 48 below.

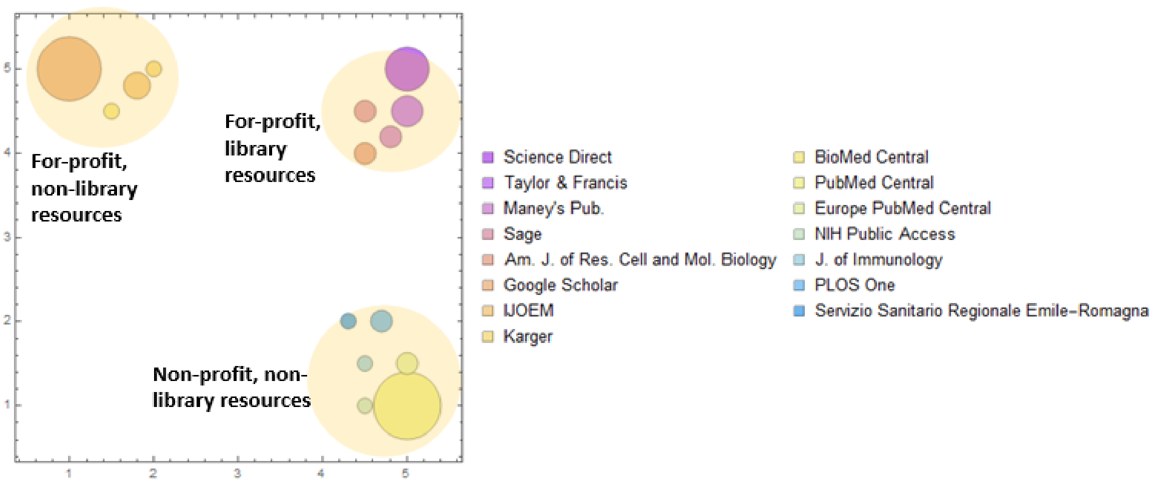
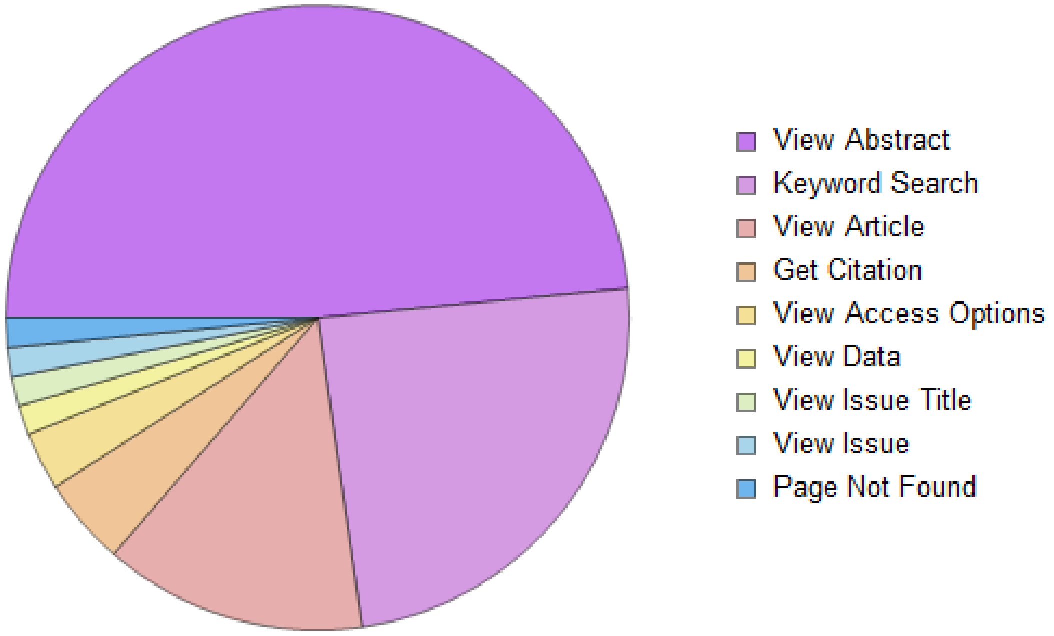


Figure 19: Dolly Grant Screenshot Data Overview

Table 43: Dolly Grant Screenshot Activities.



Activity	*Occurrence*	*Source(s)*	*Examples*
View Abstract	32	• Science Direct • Europe PubMed Central (open access) • PubMed Central (open access) • Taylor & Francis • *IJOEM: The International Journal of Occupational and Environmental Medicine* (open access but published by a commercial organization) • *The Journal of Immunology* (open access) • Karger • *American Journal of Respiratory Cell and Molecular Biology* (older issues sometimes open access) • PLOS One (open access) • Sage (open access article)	Pathological Study of Chronic Pulmonary Toxicity Induced by Intratracheally Instilled Asian Sand Dust (Kosa) Early and delayed effects of naturally occurring asbestos on serum biomarkers of inflammation and metabolism
Keyword Search	16	• Google Scholar • PubMed Central (open access)	A Biphasic Response to Silica silica and immune system
View Article	9	*All open access, including older articles in AJRCMB and Sage:* • PubMed Central BioMed Central • Europe PubMed Central • *IJOEM: The International Journal of Occupational and Environmental Medicine* • PLOS • *American Journal of Respiratory Cell and Molecular Biology* • Sage	Silica nephropathy Comparison of non-crystalline silica nanoparticles in IL-1 β release from macrophages

Activity	*Occurence*	*Source(s)*	*Examples*
Get Citation	3	• Google Scholar	Adjuvant Effect of Amorphous Silica on the Immune Response to Various Antigens in Guinea Pigs
View Access Options	2	• Maney Publishing's Online Platform • Taylor & Francis	Liver functions in silica-exposed workers in Egypt: possible role of matric remodeling and immunological factors
View Data	1	• *IJOEM: The International Journal of Occupational and Environmental Medicine*	Silica nephropathy
View Issue Title	1	• Science Direct (unauthenticated)	Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis
View Issue	1	• Science Direct (unauthenticated)	Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis
Page Not Found	1	• Servizio Sanitario Regionale Emilia-Romagna	n/a

4.6 Summary of Data Description

In this chapter, I presented my data gathering techniques in relation to multi-sited fieldwork, summarizing fieldwork and screenshot data. I described the data preparation and analysis process, which involved transcription followed by coding. I reported how I handled participant data, secondary data, and information resource attributes—including data storage and protection techniques. I developed the following code families during data analysis: interaction activities, which describe non-information-related behavior fieldwork with research participants; and information-related behaviors, which included identifying resources participants they use within the GNAE according to their attributes in relation to my primary research questions as well as—for screenshot data—activities performed by participants. This analysis enabled me to analyze data across participants for both data categories and to richly describe data for each participant in the study. I interpret these findings in Chapter Five.

Chapter 5 Findings and Discussion

5.1 Summary

In this chapter, I will interpret data gathered during the study in relation to my research questions as well as theory. I additionally review possible limitations of the study and implications for future research.

5.2 Research Questions and Findings

5.2.1 RQ1 and RQ2

In this section, I will discuss research findings in relation to the primary research questions for this dissertation: *According to the multi-sited ethnographic analysis of scientists participating in this study—individuals conducting research in various disciplines at different institutions in several geographical locations—is there evidence indicating a significant allotment of non-institutional/informal information-related exploration and discovery occurring beyond official library-supported mechanisms in the GNAE*? (RQ1) and RQ2: *What (if any) patterns are exhibited and how do these patterns relate to IS and other social science theories*?

Answering these questions requires breaking RQ1 into its constituent parts and addressing RQ2 in relation to these components. Firstly, I will summarize exploration and discovery identified during fieldwork and in screenshots which can be classified as not requiring library intermediation (i.e., beyond the paywall). I will then discuss the concept of institutional resource providers versus non-institutional ones, and finally I will introduce the concept of "informal" exploration and discovery in relation to the concept of authentication and official information versus non-authenticated, unofficial information.

5.2.1.1 Exploration and Discovery beyond the Library Paywall

Of the one hundred fifteen resources identified in both fieldwork and screenshot data, I coded eighty-eight as not being intermediated by a university library (twenty-seven being library-mediated). Twenty-six of the non-common resources originated from screenshot data, and fifty-three from fieldwork.

Figure 20 below provide visualizations for both datasets mapped to the four primary information resource attributes (yellow nodes); participants (red nodes); and the number of mentions per participant (fieldwork) or number of screenshots provided by participants in each category (edges between nodes). I have included background information regarding production of these visualizations in Appendix G.

Labels for the yellow nodes (i.e., models of production for information resources):

- FP-L: For-profit, non-library resource
- NP-L: Non-profit, library resources
- FP-NL: For-profit, non-library resources
- NP-NL: Non-profit, non-library resources

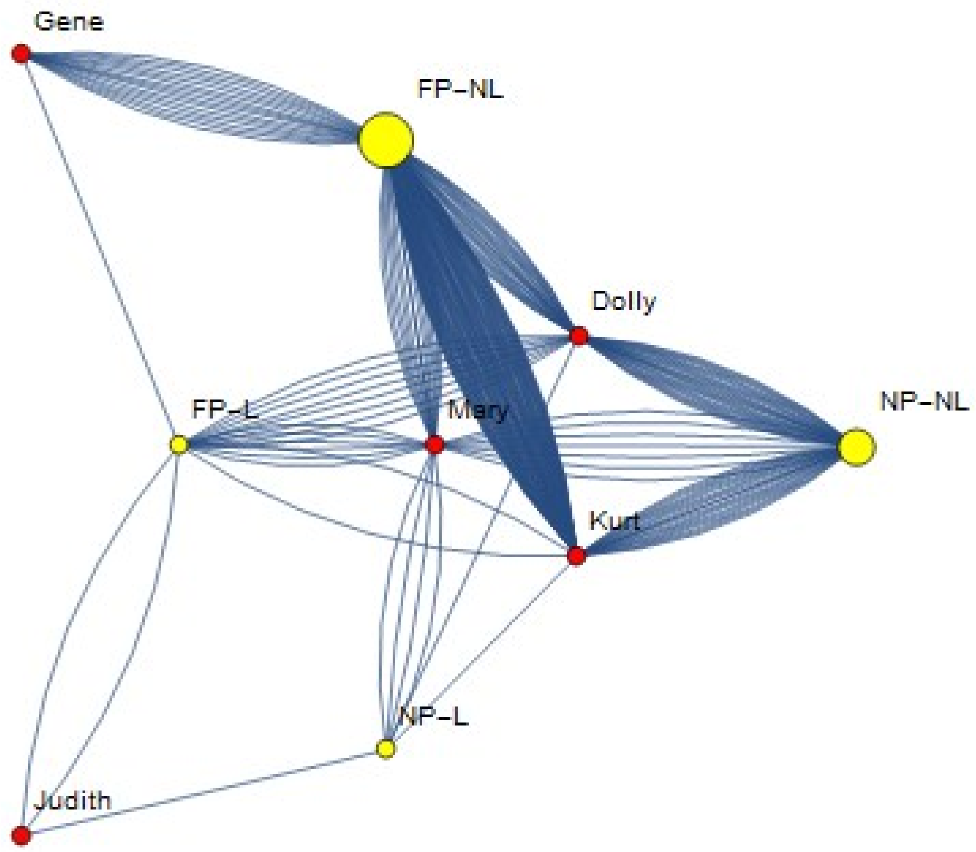


Figure 20: Combined Fieldwork and Screenshot Data: Library Versus Non-Library Resources.

For the sake of comparison, Figure 21 illustrates the scholars interacting with these same kinds of resources according to what they reported in *fieldwork conversations only*.

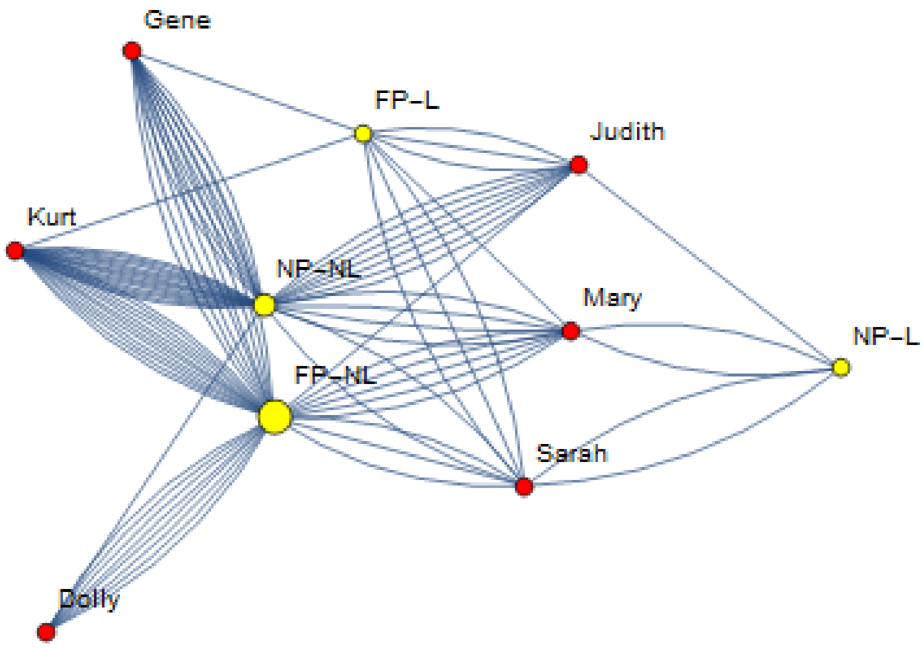


Figure 21: Fieldwork Data: Library Versus Non-Library Resources.

Both visualizations of the two datasets indicate more activity with resources not intermediated by a library (i.e., beyond library paywalls) for the six scientists in this study than those requiring a library subscription or infrastructure for delivery within the GNAE, with physicists (both theoretical and experimental) reporting in fieldwork more use of library resources than the cheminformaticians or the immunologist. Screenshot data illustrated less utilization of library-mediated resources than was reported during fieldwork, with only Judith's set of three screenshots showing exclusively library-mediated resource use.

For resources mentioned in both fieldwork and screenshots (see Section 4.5.3.3), only five required library-intermediation at an institutional level (i.e., local institution or by a university library [for arXiv, hosted by the Cornell University Library]). I included PubMed Central the non-library category here because of its national governmental support. In the following visualization, information resources not intermediated by library are yellow nodes; participants, red nodes; and library-intermediated resources, green nodes. I have combined the three Google resources mentioned in Figure 8 (Google Patent, Google Scholar, and Google CZ) into one Google node below, which is why there are twelve nodes and not fourteen.

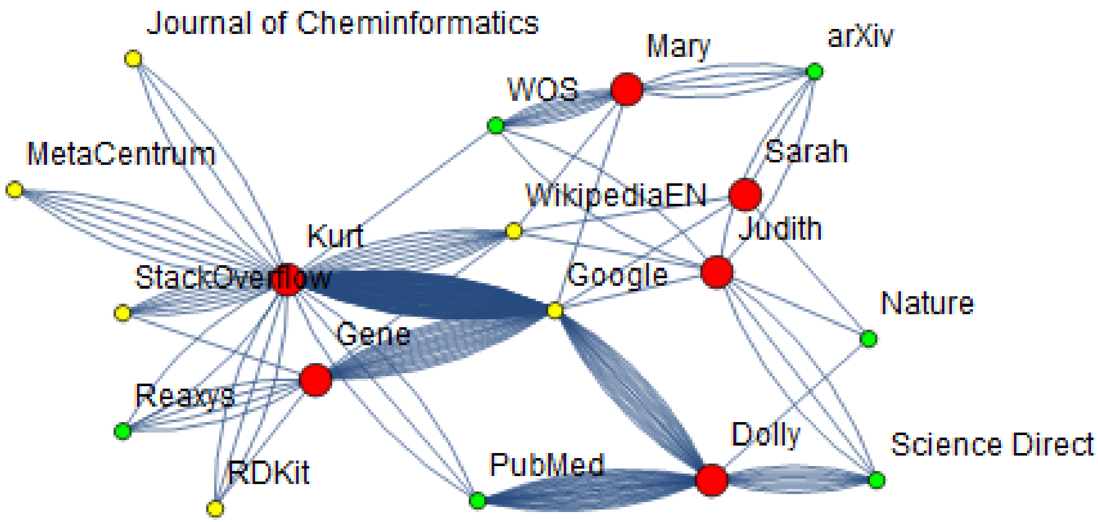


Figure 22: Resources Mentioned in Both Screenshot and Fieldwork Data: Green Nodes, Intermediated by a University Library.

Because depth of data gathered varied from participant to participant, these patterns cannot be generalized beyond the set of six participants in this study, but—for these participants—the data and visualizations thereof indicate significant information-related activity taking place beyond library paywalls. I will discuss these patterns in more detail as well as the question of *informal, non-institutional* activity in the following sections.

5.2.1.1.1 Beyond the Paywall Pattern One: Importance of Google

Resources mentioned in fieldwork and illustrated through screenshot data include seven Google products:

Table 44: Google Products.

Product	*Mentions in Fieldwork*	*Screenshots*	*Total*
Google, Czech version	1	88	89
Google Drive	1	0	1
Google Hangout	1	0	1
Google Patents	2	1	3
Google Scholar	5	18	23
Google US	0	12	12
Google (version unclear, likely Czech)	1	0	1

Google's tools are clearly where four of six participants, as illustrated by screenshot data, start exploration within the GNAE—including for article searches, which is a markedly different pattern than found by Jamali and Asadi (2010), who conclude an article about Google use by stating that:

Currently scientists do not intentionally use Google to search for articles, although this seems to be changing as they become more aware of the inclusion of Google in search results. Although at the time of data collection for this study Google was not yet very popular, this may have changed since and scholars may now have turned their attention to Google Scholar for finding articles instead of Google's general search engine (p. 291).

This study, as in another work by Boukacem-Zeghmouri and Schöpfel (2013, p. 145), found to the contrary that the two most experienced scholars at mid-career, Mary Newton and Dolly Grant, used either Google US or Google Scholar (GS) specifically to locate articles using keyword searches. Mary did use a library-intermediated tool, WOS, as well as commercial SciRate.com, which repackages arXiv content, to conduct sophisticated author searches, but Google US was her starting point for the topical searches she conducted in this study. Dolly complemented GS with abstract views in PubMed Central, but she returned again and again to GS to locate new information and to refine keyword searches. Neither Mary nor Dolly visited their library webpages during screenshot-taking sessions, although they were aware of them according to fieldwork discussion. I even asked Dolly about this after reviewing her screenshots (S. Krueger 2015, pers. corr., p6, 10 Aug):

[SK] Do you ever use the [x] library's [list of journals] **A to Z**? Why or why not?
[Dolly] No. [My] library is connected to my google scholar searches and brings up availability automatically.

Here is an illustration of keyword and query refinement progressions (in order of screenshot appearance) in Google US and GS for Mary and Dolly.

Table 45: Google Progressions.

Mary (Google US)	*Dolly (Google Scholar)*
"spray-rigid" rigidity square lattice with fixed boundary conditions robert connelly isostaticity anisotropy rigidity rigidity percolation continuous transition rigidity percolation nematic elastomer central limit theorem for products of random matrices levy localization of light	silica and immune system environment silica and immune systems Early and delayed effects of naturally occurring asbestos on serum biomarkers of inflammation and metabolism *(four different pages of search results on this topic, each with one screenshot)* Lymphocyte activation in silica-exposed workers silica and srbc A Biphasic Response to Silica *(three screenshots are different pages of results)*

[Mary's Google Progressions, Table 45 continued]	
how sorbitol regulates turgor pressure	
yeast cell turgor pressure varies inside cell	
nematic elastomers rigidity percolation	
physics venus fly trap motion	

Eight years after Hemminger et al. (2007, p. 2210) wrote that "the distinction between bibliographic/citation database and Web search engine is blurring" according to a survey of nine hundred two American basic and medical scientists, the data gathered from these participants indicate that this distinction may not be blurring, but rather scholars may be using *whatever tool works best for the task at hand from a set of resources with which they are familiar*. For example, in Mary and Dolly's datasets one observes 1) Google variants are the first stop for keyword/topical searches; 2) when Mary required sophisticated author searches, she went to WOS; and 3) Dolly used PMC to harness the power of its abstracting as well as its links to open access full-texts.

Google's various tools, for participants in this study, were not only useful for finding articles, but they also display—as Jamali and Asadi (2010, p. 288) state— "usefulness for problem-specific information seeking" as illustrated by the screenshots provided by Gene Kim and Kurt DeSilva. For both Gene and Kurt, Google functioned as a handy, ever-present tool for asking questions about many topics. Kurt commented on this phenomenon in relation to his screenshots (a selection of his comments follows below; see Appendix G for all commentaries). Gene's queries focused mainly on programming questions, with Kurt's displaying a wider variety of topical interest. As noted in the previous chapter, although Kurt and Gene are Czech, their day-to-day work with information resources in the GNAE is conducted in English. I have included the full set of Kurt's screenshot comments in Appendix H.

Table 46: Ask Google and It Answers

Kurt (Google CZ); selected Google screenshot queries and commentaries	*Gene (Google CZ); all Google screenshot queries*
metathesis ruthenium removal *[KdS] I had a pharma assignment to find an industrially viable way of producing the 'rolapitant' drug from available resources. A part of rolapitant making process is olefin metathesis, which uses ruthenium-based catalysts. Ruthenium is quite toxic, so I was searching if I can get it out of the reaction mixture after it has done its job.*	python postgresql postgre show tables postgre limit offset

[Table 46, continued: Kurt (Google CZ); selected Google screenshot queries and commentaries] Hoveyda-Grubbs Catalyst *[KdS]A search for the catalyst in question. Is it available?* US 20140031549 A1 *[KdS]The original patent describing, among many other things, a working (but just lab-scale) rolapitant-making process. This was the most informative source by far, but also a bit annoying to read due to referencing many other papers…* celite pad *[KdS] I wasn't sure what the 'celite pad' mentioned in the patent was, so I googled it and facepalmed the moment I saw the pictures. Yes, I am a theorist and it really shows :)* Brine *[KdS] The same as above, just search for 'brine'. Turns out it's just a fancy word for salty water.* pd c hydrogenation *[KdS] Another step in rolaptitant prep is hydrogenation catalyzed by Pd on C. I was already somewhat familiar with this, but googled just to make sure.* lithium tris(tert-butoxy)aluminium hybrid *[KdS] Was searching for this compound mentioned in the patent.* latex figure fixed position *[KdS] I needed to place an unruly figure in my rolapitant report that I was typesetting with LaTeX.* chiralcel separation *[KdS] Didn't know what 'ciracel separation is. Seems to be a proprietary lab method of separating isomers.* fluoro group substitution *[KdS] Can't remember why I searched for that.* linux generate md5 checksum *[KdS] Generic linux stuff, about generating checksums to validate my program exports, I think.* numpy metacentrum *[KdS] Searched whether MetaCentrum where I compute most of my stuff supports Numpy (a python library for advanced computational/statistical things* django haystack searchqueryset *[KdS]Django is a Python framework that I use to provide some sort of user interaction and handling more complex processing over my chemical database. Now I was search[ing] for a way to reset the database (to clear it), but not to drop/delete the tables themselves.*	*[Gene (Google CZ); all Google screenshot queries]* postgre insert info brufen wiki create alias bash postgres count connections drop column postgres postgres left outer join python regex sql select random rows r quick start guide r # csv postgres reset sequence python exception

This pragmatic behavior of asking Google for answers indicates that for participants in this study—particularly Gene and Kurt—Google functions as a "surrogate expert" (Souto-Otero and Beneito-Montagu 2013 citing Simpson 2012, p. 11). However, all participants in this study did note in fieldwork conversations that they do ask colleagues for information as well, and often this communication takes place in-person (locally) and over email (locally and globally), as found by Allard et al. 2009 and Levin et al. 2011. Kurt, for example, noted (2014, S. Krueger pers.comm., 17 Aug):

> In my limited experience, most…collaborations are handled by personal/private communication methods like mail, phone, skype, private meetings, etc. With [x] and [x], we discuss vast majority of things via mail or personally in meetings, both periodical and problem-specific. This doesn't leave much of a trace for anyone outside, I'm afraid (well, except perhaps the NSA :)

5.2.1.1.2 Beyond the Paywall Pattern Two: Importance of Wikipedia (English)

Four participants mentioned in fieldwork the importance of English Wikipedia for exploration and discovery. It has become, for participants in this study, the *de facto* replacement for specialized handbooks or encyclopedias, and—for the three of these participants not based in the US, it is trusted more than their regional language versions.

No participants mentioned that they edit Wikipedia entries themselves, and in terms of the scholarly context regarding Wikipedia, "little is known about editors of science-related articles" (Teplitskiy et al. 2015, p. 4 citing West et al. 2012). Regarding articles in chemistry, however, Van Noorden (2012) mentions a study by Andrew Williams (creator of ChemSpider, a free chemical structure database) that found chemical structures in Wikipedia are highly accurate due to efforts by chemists to "clean up" such articles themselves. I was unable to locate any prior research specifically on the topic of the quality of Wikipedia articles in condensed matter physics or immunology/toxicology.

Table 47: Wikipedia and Participant Commentary Summary.

Product	*Fieldwork Commentaries*	*Screenshots*
Mary	"Daily use"	0
Judith	"Is not very scientific" but "uses for definitions" and "the English scientific articles are much richer than those in the [her] native language"	0
Kurt	"I use wikipedia to look up things that I know very little about"	7 – Brine; Sulfonamide (medicine); Citation from *The Lancet*; Cocrystal – two screenshots; Transcription factor; Glucocorticoid receptor
Gene	Uses "for general information, introduction to a topic"	0

This pattern of Wikipedia use by four participants coincides with findings from Aibar et al. (2015, p. 8), whose survey of nine-hundred thirteen university faculty members in Spain found that "faculty use Wikipedia for consulting matters not strictly belonging to their field of expertise" and that, unlike earlier studies on Wikipedia use, "faculty actually think the quality of Wikipedia articles is quite respectable" and while articles could be more comprehensive, "they are clearly considered reliable and updated" (p. 7). I will discuss trust in Wikipedia in addressing the question of informal, non-institutional use as an example of a mass-online "commons-based peer production" (Benkler 2006, p. 60) in Section 5.2.1.2 below.

The importance of Google and Wikipedia for participants in this study raises many interesting questions for future researchers about how Google and Wikipedia are entwined (or not) in the creation of Google's Knowledge Vault (Dong et al. 2014). Simonite (2013) comments superficially on Wikipedia's contribution to the Google fact box, but how much it contributes to Knowledge Vault or what specific role it plays in Google's Knowledge-Based Trust (KBT) algorithm (Dong et al. 2015), which currently being tested for addressing the "correctness of information provided by a web source" (p. 12), is unknown as of May 2016. Exploration of such questions is, however, beyond the scope of this dissertation.

5.2.1.1.3 Beyond the Paywall Pattern Three: Open Resources and Tools and the Disintermediation of the Library

Both fieldwork and screenshot data illustrate the importance of open resources for the six scientists in this study for exploration and discovery in the GNAE, as seen in the data description for each participant (Section 4.5.4). Open resources and tools mentioned or observed in screenshots include not only open access journals (e.g., *The Journal of Cheminformatics*), but also open software or programming tools (Python, RDKit) and other platforms and databases created by governments (PubMed, PubChem) or institutions (ChEMBL) for use by specialized scientific communities.

The first notable pattern here is the *total disintermediation of the library* from the exploration and discovery process for participants who start their searches in Google variants off-campus without authenticating to a library proxy server. In other words, *the library paywall becomes a completely invisible and insurmountable barrier* for users who are not aware of institutional remote access options for accessing content to which their libraries actually subscribe. In such cases, as illustrated by both Mary and Dolly, a user has several choices, none of which include libraries as intermediaries:

1) Purchase the item directly from the publisher; in this study, only Sarah reported this behavior—she "sometimes ends up purchasing books" herself (S. Krueger 2013, Fieldwork Transcriptions, p. 10, 3 Nov).

2) Try to find a pre-print or other form of open—even "illegally" open—content; in this study, both Mary and Dolly did this, as seen directly below.

3) Give up on reaching a full-text and move on to another item, as seen in Mary's *Nature* example (see Section 4.5.4.2.2).

Regarding the second choice, it is useful to analyze screenshot data in order to better understand the phenomenon.

In the following screenshot, Dolly begins her search in Google Scholar, without library links enabled. She does not notice the links settings are not working, and thinks she is searching over library content (DG: [My] library is connected to my google scholar searches and brings up availability automatically. [S. Krueger 2015, pers. corr., p6, 10 Aug]). Only one full-text PDF appears to be available to her.

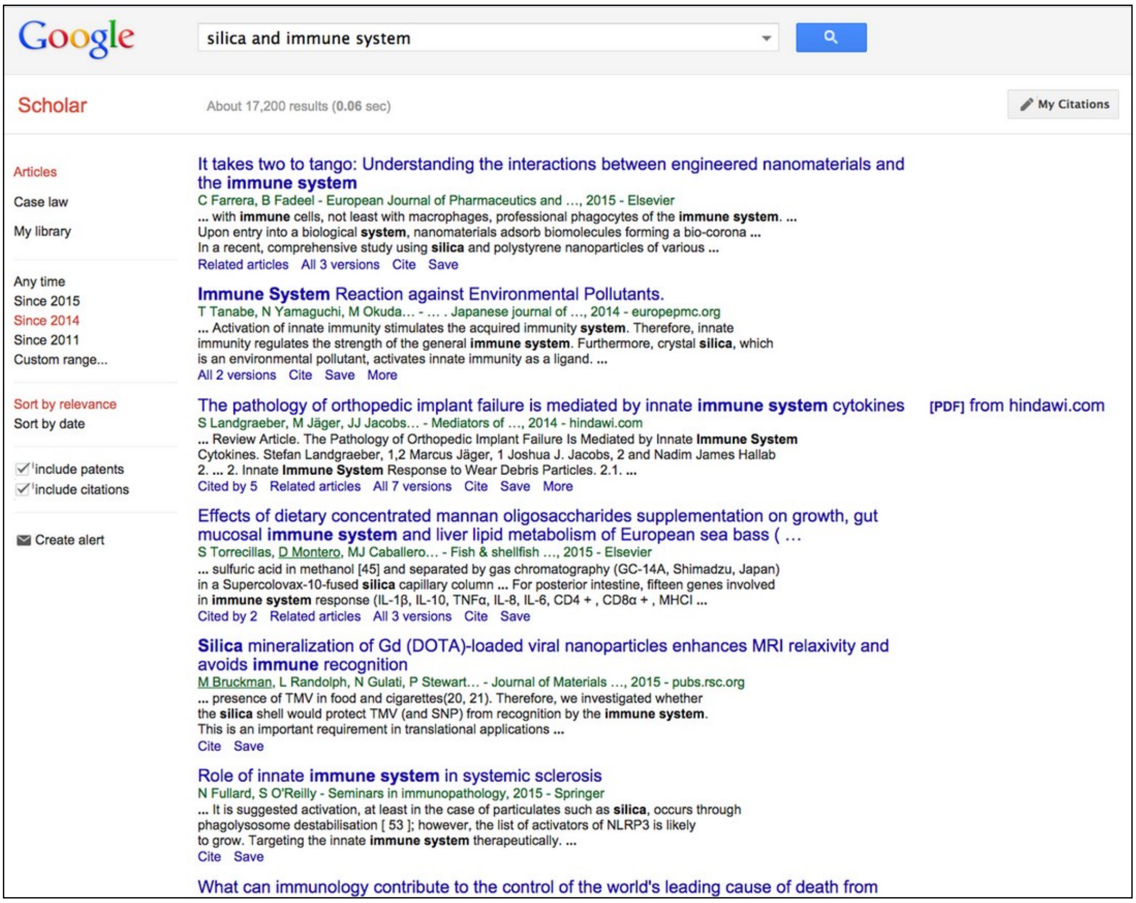


Figure 23: Google Scholar without Library Links Enabled (D. Grant 2015, screenshot data, 6 Aug).

Had Dolly's library links in GS been enabled properly, links to additional full-text available inside her library's paywall would have appeared. I tested activating links to Dolly's library, and the results list changes dramatically, with several full-texts provided by her library available within two clicks. In Figure 24 below, I have blacked out the name of Dolly's institution and left my library settings (NTK) to illustrate GS's multiple library linking capabilities. Currently, only five library link settings are allowed in GS.

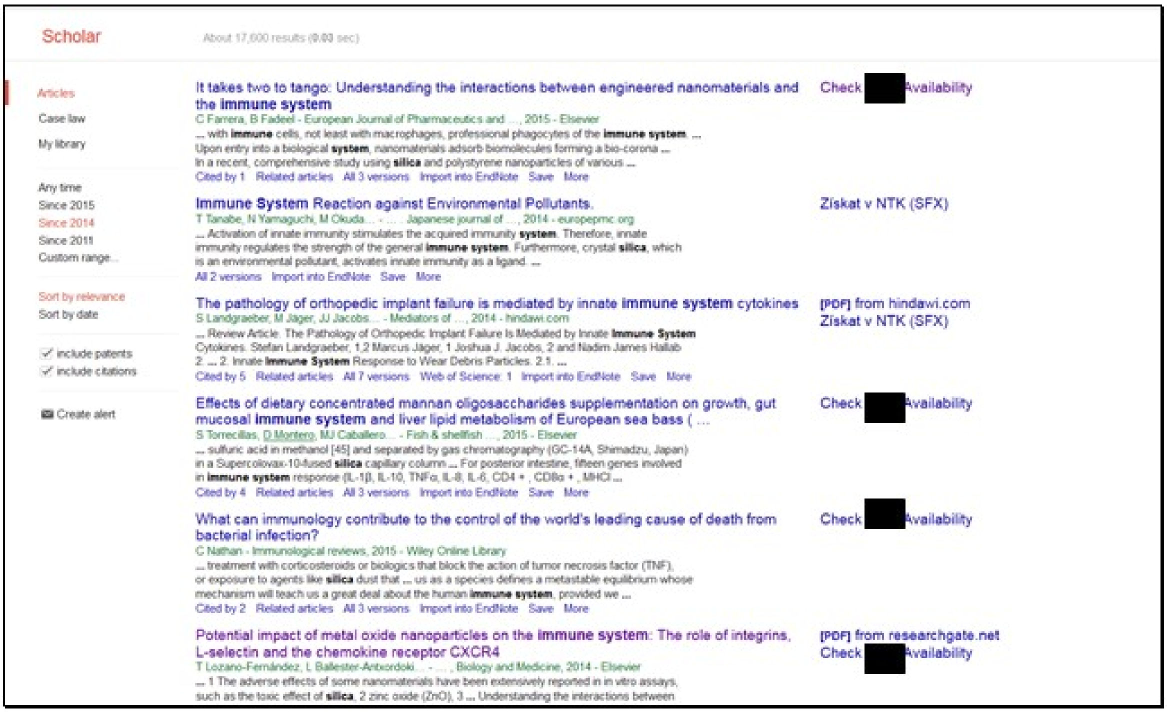


Figure 24: Google Scholar with Library Links Enabled (S. Krueger 2015, screenshot, 25 Aug).

Because her settings were not working properly, Dolly selected the first item by clicking on the GS link (*It takes two to tango)* and encountered the Elsevier paywall. After viewing the article abstract, she moved on to other items, utilizing *only open access full-texts* for the duration of her searching session illustrated by screenshots.

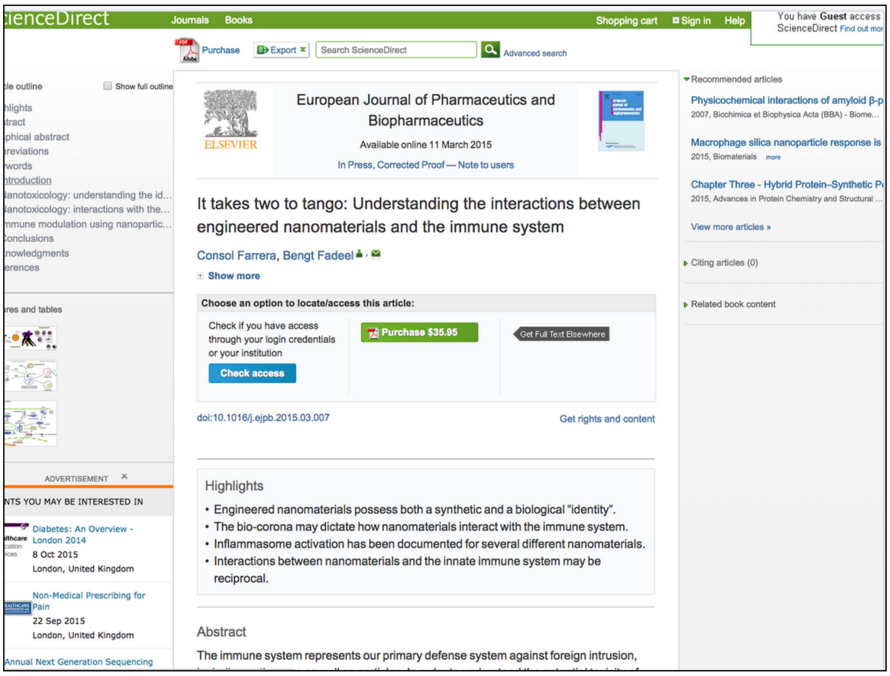


Figure 25: Elsevier Paywall (D. Grant 2015, screenshot data, 6 Aug).

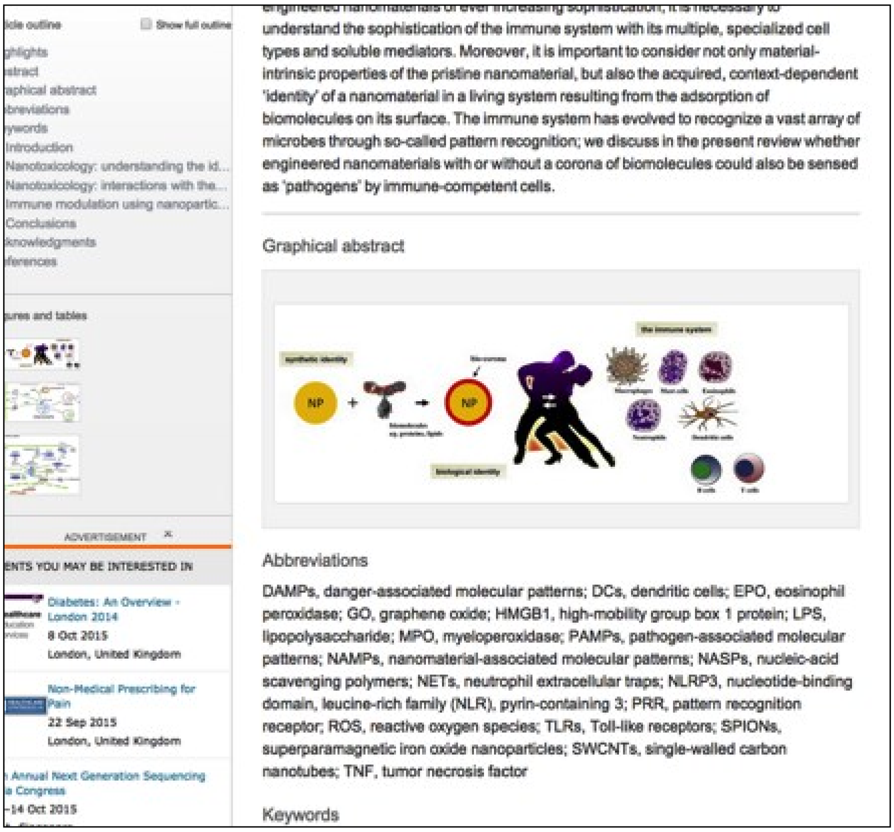


Figure 26: Viewing Abstract (D. Grant 2015, screenshot data, 6 Aug).

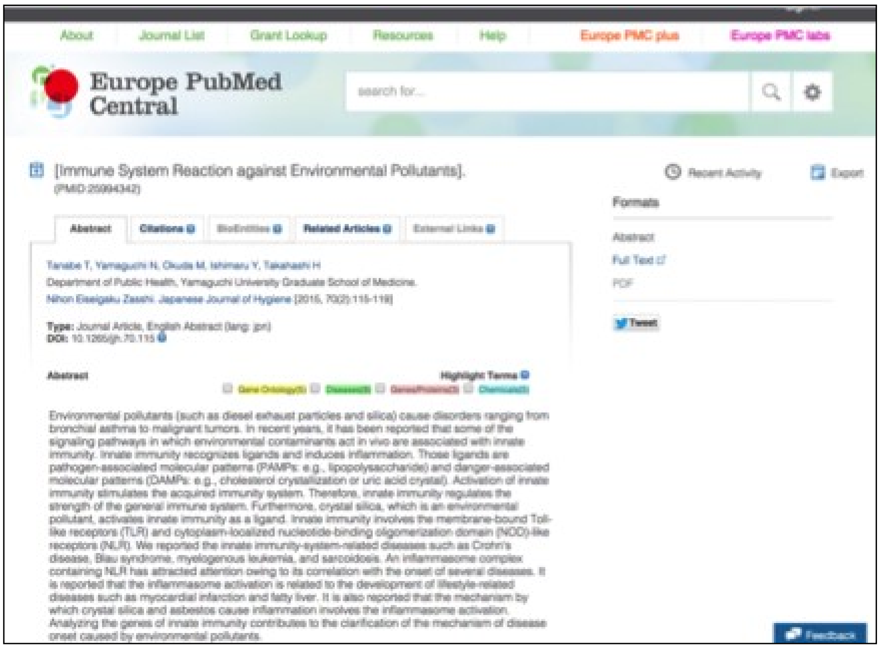


Figure 27: Open Access Full-Text of Another Article at Europe PubMed Central (D. Grant 2015, screenshot data, 6 Aug).

Mary exhibits a similar pattern of search beyond the library paywall, using Google instead of GS, which enabled her to find a broader spectrum of PDFs and full-texts made available by professors themselves on their institutional servers.

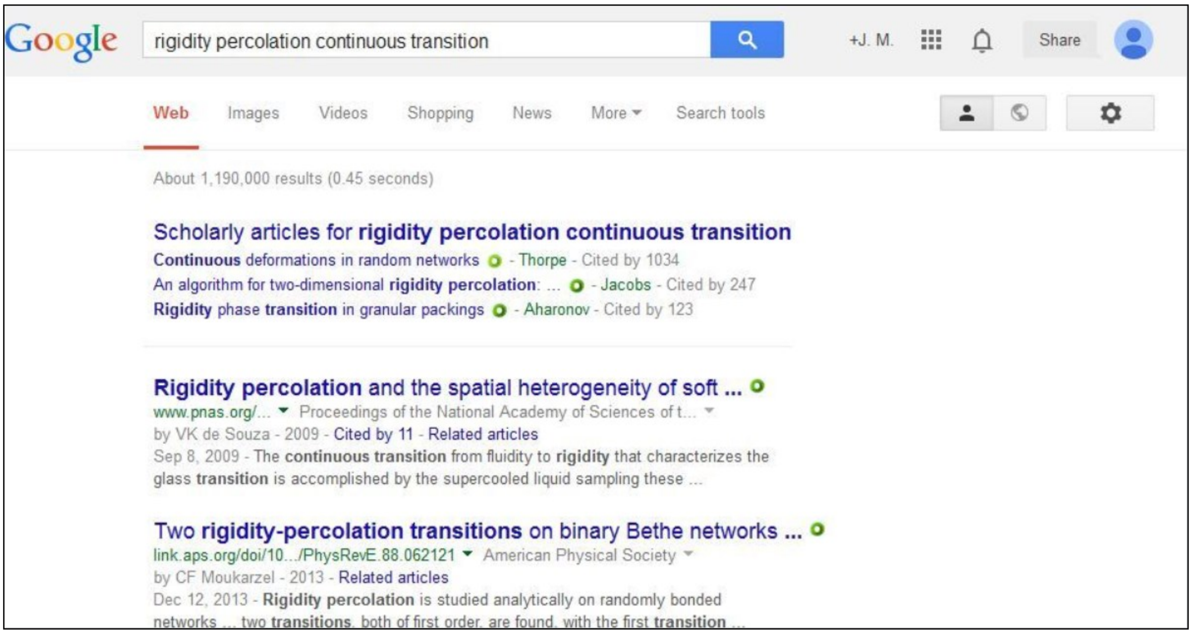


Figure 28: Google Query. (M. Newton 2014, screenshot data, 18 Aug).

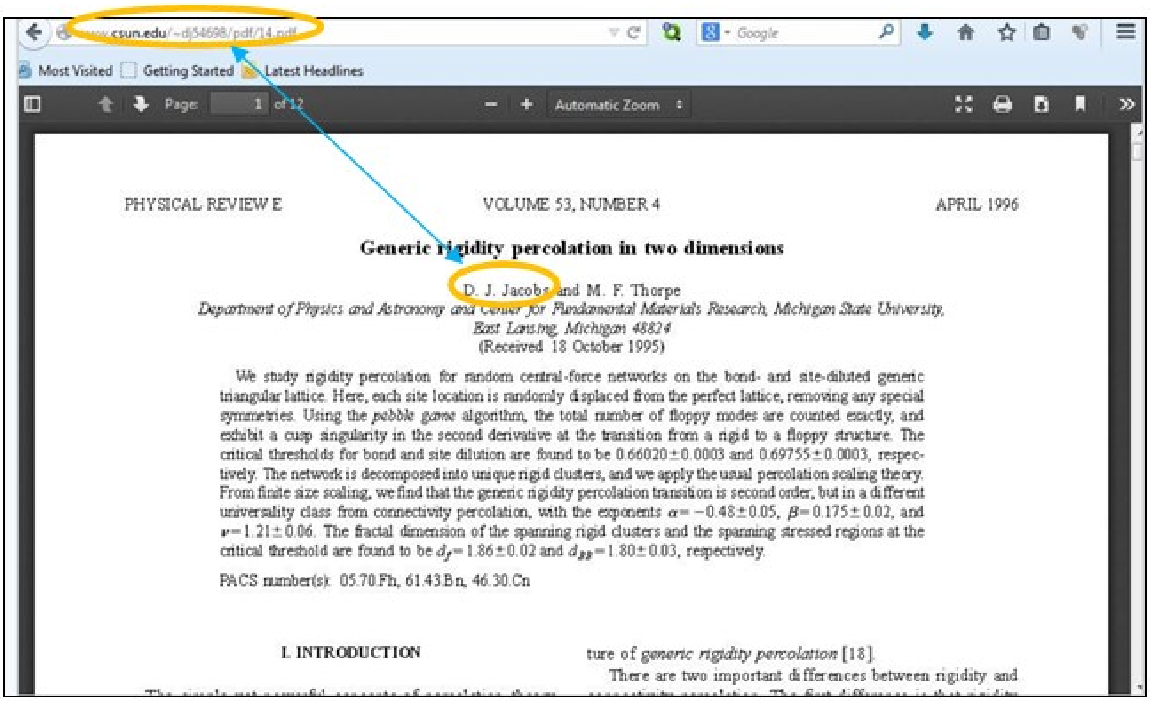


Figure 29: PDF on Author's Institutional Server Space. (M. Newton 2014, screenshot data, 18 Aug).

Using Google instead of GS enabled Mary to access a conference presentation which is not indexed in GS (as of August 2015):

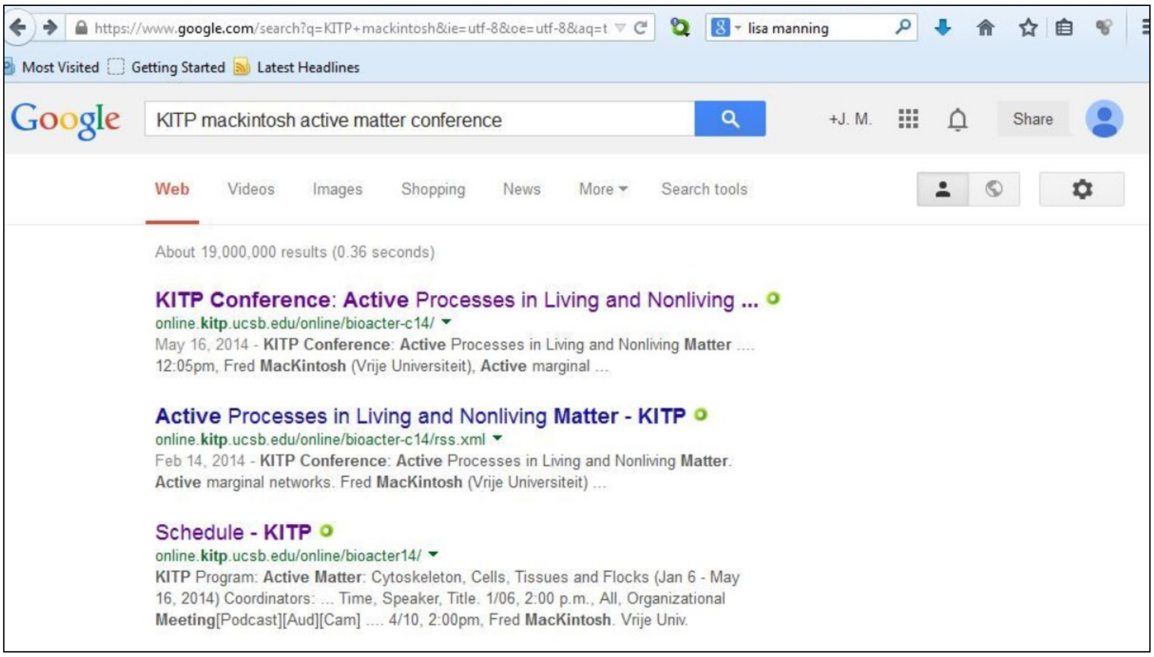


Figure 30: Using Google to Find Conference Presentation. (M. Newton 2014, screenshot data, 18 Aug).

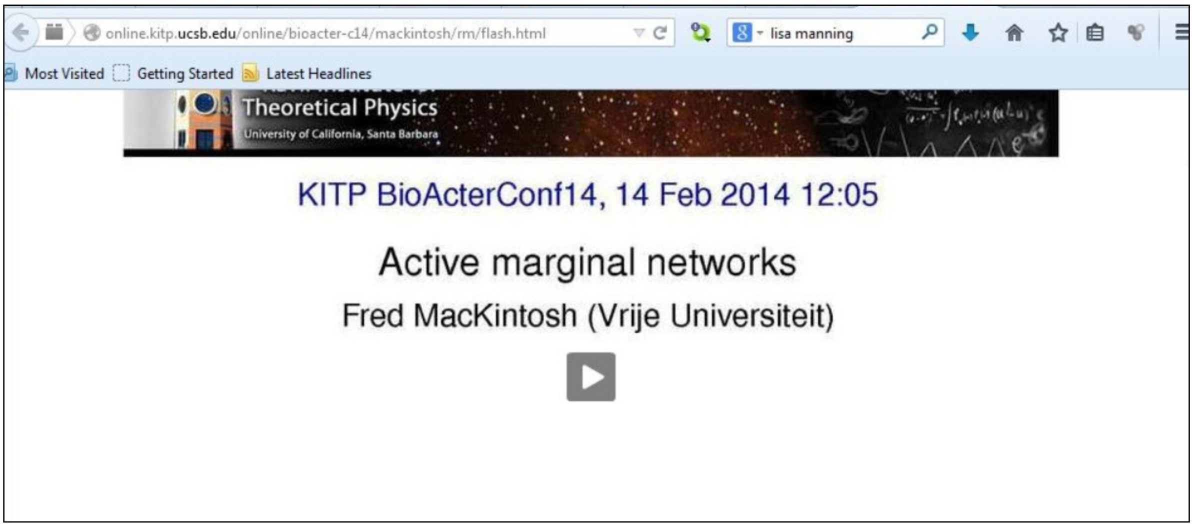


Figure 31: Viewing Conference Presentation. (M. Newton 2014, screenshot data, 18 Aug).

While no participants provided screenshot data illustrating use of "illegally" uploaded content, several European participants as well as doctoral students I have met while embedded on the Dejvice technology campus have commented that they commonly use of filesharing servers and/or websites. I asked Kurt about this for Czech content, and he noted:

> Copyright protection here seems to be somewhat lacking (it's not just torrents and filesharing hubs—for example, uloz.to download service is full of proprietary material and it has been active for years) (S. Krueger 2015, pers. corr., 21 Aug).

In addition to uloz.to, another popular filesharing service on NTK's surrounding technology campus is Library Genesis (LibGen), based in Russia. Elsevier is currently trying to shut down LibGen, but with limited success—as of May 2016, LibGen was active and fully-accessible at Gen.lib.rus.ec, as noted by Ernesto (2015) and those who commented on his article in August 2015:

3 ^ | v · Reply · Share ›

thisguy1337 · 2 months ago
Get the dumps while you can...
http://gen.lib.rus.ec/dbdumps/

Mirror list from the site:
http://gen.lib.rus.ec/
http://bookfi.org/
http://scienceengineering.libr...
http://bookzz.org/
I2p : http://u76v7ha6j4jmtz3k2lseaso5qy36lxs77klhovmptufwcodovatq.b32.i2p/

DOWNLOAD UPLOAD

Mirrors

libgen.in - 1M (Main)

see more

13 ^ | v · Reply · Share ›

James Edward Lewis → thisguy1337 · 2 months ago
I just noticed you posted a link to a database dump of scientific articles a couple weeks ago, on an earlier story about LibGen, but now that dump is gone: http://gen.lib.rus.ec/dbdumps/...
^ | v · Reply · Share ›

thisguy1337 → James Edward Lewis · 2 months ago
I know =(nothing I can do about that sadly...
3 ^ | v · Reply · Share ›

DiD → thisguy1337 · 2 months ago
cough https://mega.nz/#F!EgcCClIY!4H...

database dump June 19 incl. usbwebserver for local search.
See https://genofond.org/viewtopic...

EDIT: The .ec site still returns scientific papers, but there is no more dump for them
5 ^ | v · Reply · Share ›

F. Tioli → thisguy1337 · 2 months ago
Thanks for the links! But do these dumps include the full articles, or is it just the database?

Thanks in advance
4 ^ | v · Reply · Share ›

Feniks → thisguy1337 · 2 months ago
Arigato the top mirror is still up as 8.33GMT
^ | v · Reply · Share ›

TxRx100 → thisguy1337 · 2 months ago
Why are there two dumps, http://gen.lib.rus.ec/dbdumps/ and

Figure 32: LibGen: "Get the dumps while you can." (S. Krueger 2015, screenshot, 25 Aug).

I asked Mary about illegal downloads in November 2014, and she was aware of this topic as well, which indicates this phenomenon is not limited to European fieldsites:

> Regarding illegal downloads, if you submit a paper to a journal, they never ask how/where you acquired access to the articles you cite. My policy is that if the students are downloading to a computer that I purchased with government funds, then they should not do it. However, if it is to their own computer, then be careful. (S. Krueger, 2014, pers. corr., 20 Nov)

Regardless of how material in the GNAE is made openly available beyond the library paywall (legally or illegally), data gathered in this study confirm the "amplifying impact of open access" on the "diffusion of science" in cases when the library is disintermediated, as identified by Teplitskiy et al. (2015) in relation to Wikipedia. Not only are open items (articles or items such as the conference presentation in Figure 31) the *only* visible items for users who are not authenticated to libraries through proxy servers, as illustrated in the cases of Dolly and Mary, they are also able to be repackaged and distributed in new ways, such as in Wikipedia articles, which are in turn—as illustrated by this study—utilized by scientists (and others) around the globe:

> Editors of the English Wikipedia act as 'bootleggers' of high quality science by distilling and distributing otherwise closed access knowledge to a broad public audience, free of charge. Moreover, the English Wikipedia, as a platform, acts as an 'amplifier' for the (already freely available) open access literature by preferentially broadcasting its findings to millions. (Teplitskiy et al. 2015, pp. 24-25)

The "amplification effect" of open access would be an interesting point of departure for future research.

Regarding open resources and tools themselves and their models of production, this study found that there is not yet one model which is preferred; the six scientists *use open resources and tools pragmatically and in a manner which fits their research questions at hand*. No participants mentioned that they contemplate the for- or non-profit status of a journal or content producer except for Gene, who noted that while many of the chemical databases he uses are commercial, that he does "try to use open ones" (S. Krueger 2014, Fieldwork Transcriptions, p. 23, 9 Jul). That being case, there are two notable patterns to be gleaned from this study: 1) for traditional scholarly journals, *open access mandates (by governments or funding foundations) make a notable difference* in expanding the global accessibility of scholarly content, as observed in Dolly's work with resources such as PubMed Central and European PubMed Central, and 2) for scientists in this study, *fewer than half of* *the resources* mentioned or illustrated being used in screenshots originate from for-profit providers (seventeen of the forty-five in the table below). While commercial publishers remain relevant for participants in this study, they are supplemented with a wide array of other materials made available in the GNAE by non-commercial players.

The following table summarizes open resources and tools (including freemium tools and excluding research groups, mailing lists, and Google variants) mentioned in fieldwork and screenshots which are not facilitated via a local/university library infrastructure. I introduce the concepts of institutional or non-intuitional (i.e., peer-community) creation, which I will address in the next section in relation to RQ1's *non-institutional/informal* information-related exploration and discovery component.

Table 48: Open Resources and Tools.

Resource	*Description*	*For- or non-profit, institutional or non-institional*
All Things Metathesis	Blog	For-profit; Materia, Inc.
American Journal of Respiratory Cell and Molecular Biology *(for NIH-funded research only)*	Scholarly journal	Non-profit; American Thoracic Society
BibTeX	Document preparation	Non-profit; peer-community created
BioMed Central	Scholarly journals	For-profit; Springer Science+Business Media
BitBucket (open for up to five users)	Team code management	For-profit; Atlassian software
ChEMBL	Bioactive data for drug discovery	Non-profit; European Bioinformatics Institute/Wellcome Trust
chemicalize.org	Public resource developed by ChemAxon which uses ChemAxon's Name to Structure parsing to identify chemical structures on webpages and other text.	For-profit; ChemAxon
Dropbox (freemium model—free to limit)	Filesharing	For-profit; Dropbox, Inc.
Eclipse	Open source Integrated Developing Environments (IDE) and tools	Non-profit; Eclipse Foundation
Electronic Communications in Probability	Scholarly journal	Non-profit; Institute of Mathematical Statistics with Bernoulli Society
Espacenet	Patent search, European Union	Non-profit; European Patent Office
European PubMed Central	Free biomedical research literature	Non-profit; Europe PMC Consortium
FDNMES	Code to simulate X-ray spectroscopies	Non-profit; Institut NEEL Grenoble
FullProf Suite	Crystallographic programs for	Non-profit; Institut Laue-

[Table continued: Resource]	Rietveld analysis (structure profile refinement) of neutron *[Description]* (constant wavelength, time of flight, nuclear and magnetic scattering) or X-ray powder diffraction data collected at constant or variable step in scattering angle 2theta.	Langevin Grenoble *[For- or non-profit, institutional or non-institional]*
GitHub (freemium—free for public repositories)	Coding tools, tracking, team coding	For-profit; GitHub, Inc.
The International Journal of Occupational and Environmental Medicine	Scholarly journal	For-profit; NIOC Health Organization, part of the National Iranian Oil Company
JabRef	Open source bibliographic reference manager	Non-profit; peer-community created
Java (GNU public licensed components)	Programming language	For-profit; Oracle Corportation
Journal of Cheminformatics	Scholarly journal	For-profit; BioMed Central Ltd.
Journal of Immunology (open access from 1998 to one year prior to current issue date)	Scholarly journal	Non-profit; American Association of Immunologists
Journal of Investigative Dermatology (open access for research funded by the US National Institutes of Health, Howard Hughes Medical Institute, and the Wellcome Trust)	Scholarly journal	For-profit; Nature Publishing Group
Journal of Medicinal Chemistry (open access for some research funders plus at author's discretion)	Scholarly journal	Non-profit; American Chemical Society
Karger (selected historical journals via open access portal)	Scholarly journals and books	For-profit; Karger Publishers
KNIME	Open source data analytics, reporting, and integration	Non-profit; peer-community created (originally: University of Konstanz)
LaTeX	Document preparation	Non-profit; peer-community created
LinkedIn (freemium)	Profile creation	For-profit; LinkedIn, Inc.
Maney's Publishing (open access for some research funders plus at author's discretion; selected journals after one year)	Scholarly journals	For-profit; Maney's Publishing
OCHEM (Online chemical database with modeling environment)	Chemistry database with modeling and predictions	Non-profit; federation of supporters, both for- and non-profit (ChemAxon, Chemistry Development Toolkit, Molecular Networks GmbH, TALETE, and InChi Trust)
Open Babel	Open source chemistry toolbox	Non-profit; peer-community created
PLOS (Public Library of Science)	Scholarly journals	Non-profit; PLOS

[Table continued: Resource] PNAS (Proceedings of the National Academy of Sciences; open access after six months, with some immediate access)	*[Description]* Scholarly journal	*[For- or non-profit, institutional or non-institutional]* Non-profit; US National Academy of Sciences
PubChem	Database of chemical molecules and their activities against biological assays	Non-profit; National Center for Biotechnology Information, part of the National Library of Medicine, US National Institutes of Health
PubMed	Free tool for accessing MEDLINE content with links to full-text, when available, in PUBMED Central and its local mirrors	Non-profit; National Library of Medicine, US National Institutes of Health
Python	Programming language	Non-profit; Python Software Foundation (originally peer-community created)
RDKit	Open source cheminformatics software	Non-profit; peer-community created
SAGE (limited gold and hybrid gold open access)	Scholarly journals	For-profit; SAGE publishers
SciPy	Open source software for scientific computing in Python	Non-profit; peer-community created
SciRate	Scientific collaboration network	For-profit (.com); peer-community created but cannibalizes arXiv content
Skype (freemium)	Telecommunications software	For-profit; Skype Technologies
Stack Overflow	Questions and answers for professional and enthusiast programmers	For-profit; Stack Exchange, Inc.
Surface Evolver	Interactive program for the modelling of liquid surfaces shaped by various forces and constraints	Non-profit; Susquehanna University (Pennsylvania, US)
Taylor & Francis (limited open access)	Scholarly journals	For-profit; Taylor & Francis Group
UDock	Free interactive protein docking system	Non-profit; Conservatoire National des Arts et Metiers, France, by the Centre d'Etudes et de Recherche en Informatique et Communications (CEDRIC) and the Laboratoire de Genomique, Bioinformatique, et Applications (GBA)
Wikipedia English and Ukranian	Collaborative encyclopedia	Non-profit; Wikipedia Foundation/peer-community created
Xmgrace	2D plotting tool	Non-profit; peer-community created

5.2.1.2 Informal and Non-Institutional Exploration and Discovery Patterns

In this section, I will address RQ1's non-institutional, information information-related exploration and discovery component, having discussed general patterns of activity beyond library paywalls above.

Unraveling this component of RQ1 involves introducing the concept of what Benkler (2006) has identified as a "new model of production" (p. 59), one that is "decentralized, collaborative, and nonproprietary…[in other words,] commons-based peer-production" (p. 60). Benkler states that this model evolved side-by-side with the networked environment itself as well as the open source software (OSS) movement:

> The networked environment not only provides a more effective platform for action to nonprofit organizations that organize actions like firms or to hobbyists who merely coexist coordinately. It also provides a platform for new mechanisms for widely dispersed agents to adopt radically decentralized cooperation strategies other than by using proprietary and contractual claims to elicit prices or impose managerial commands. This kind of information production by agents operating on a decentralized, nonproprietary model is not completely new. Science is built by many people contributing incrementally—not operating on market signals, not being handed their research marching orders by a boss—independently deciding what to research, bringing their collaboration together, and creating science. What we see in the networked information economy is a dramatic increase in the importance and the centrality of information produced in this way. (Benkler 2006, p. 63)

For the scientists who participated in this study, while formal scholarly journals and traditional peer-reviewed resources are still utilized and represent the dominant model of scholarly publishing and research activity, there is evidence of a notable shift in the *models of production* behind some of the open resources used beyond library paywalls—specifically, open source tools for document preparation, data analysis and programming as well as in the peer-community produced aspects of dominant non-library resources used by the participants in this study, Google and Wikipedia. The aspects of *models of information production* and peer-community production, in particular, have been almost completely neglected in previous IS research (as of date of publication of this dissertation, only Watkinson et al. 2016 touch upon the topic, providing an illustration of "dimensions" of trust for different information sources/channels, p. 448).

For Google, Benkler (2006, p. 76) notes how the PageRank algorithm "harnesses peer production of ranking" by counting "links as distributed votes of confidence in the quality of the page pointed to," thus enabling "users to find things they want quickly and efficiently." Wikipedia's "open peer-produced model" (p. 71) which "depends on self-conscious use of open discourse, usually aimed at consensus" (p. 72). A key concept here is a *new definition of trust in the authenticity—reliability and credibility—of information resources* which have been vetted by decentralized mechanisms (i.e., Wikipedia's global network of editors or, in Google's case, an algorithm built on trust in links as votes) rather than by traditional centralized scholarly peer review conducted via organizational hierarchies.

This concept of trust in relation to authenticity provides a way for me to define "formal" versus "informal" resource. For the purposes of this study, a "formal resource" is defined as one *given the stamp of authenticity by a centralized authority or institution such as a scholarly publisher or a peer-review board*, which certifies it as an official output of the authority or institution which produced it. The concept of *authentication* here is an extension of a definition developed by the American Association of Law Libraries in relation to online legal resources (Matthews and Baish 2007, p. 20):

> An *authentic* text is one whose content has been verified by a government entity to be complete and unaltered when compared to the version approved or published by the content originator…An *authentic* text is able to be authenticated, which means that the particular text in question can be validated, ensuring that it is what it claims to be.

"Formal resources" here include open access scholarly journals articles, because the primary model of production behind these materials includes traditional peer-review. Passing the peer-review process provides such articles with the *stamp of authenticity* of the journal. The publishing of the article in a scholarly journal—gated or with open access—then affirms this authenticity of the journal's publisher.

"Informal information resources," in contrast, are ones which rely on peer-production of their content or which have little or no editorial oversight whatsoever (e.g., personal emails). Peer-production is inherently decentralized and collaborative and can take place with the support of an institution (e.g., Google as "corporate host" of the PageView algorithm or an organization such as the Wikimedia Foundation) or it can be completely decentralized without requiring institutional support except server space somewhere, as is the case with many open software and tools. I would include arXiv pre-prints in this category, because they are reviewed by a global network of moderators (Ginsparg 2011).

With this definition in place, it is possible to make a list of *informal information resources* identified in this study which do not have a traditional institutional (i.e., scholarly publishing or library-related) host and which are beyond library paywalls.

Table 49: Informal Resources Developed by Peer Communities.

Resource	*Description*	*For- or non-profit, institutional or non-institional*
BibTeX	Document preparation	Non-profit; peer-community created
JabRef	Open source bibliographic reference manager	Non-profit; peer-community created
KNIME	Open source data analytics, reporting, and integration	Non-profit; peer-community created (originally: University of Konstanz)
LaTeX	Document preparation	Non-profit; peer-community created
Open Babel	Open source chemistry toolbox	Non-profit; peer-community created
RDKit	Open source cheminformatics software	Non-profit; peer-community created
SciPy	Open source software for scientific computing in Python	Non-profit; peer-community created
Xmgrace	2D plotting tool	Non-profit; peer-community created

What remains here are the open source tools for document preparation, data analysis and programming created and made available by communities of peers working together across space and time in the GNAE. These resources have not undergone peer-review by a formal hierarchical process, but are *informally given trust* by communities of peers. Benkler (2006, p. 68) refers to trust in this context as credibility, which "is a question of quality by some objective measure that the individual adopts as appropriate for purposes of evaluating a given utterance." In other words, trust here is relative in relation to the particular peer community and can therefore exhibit a level of formality (or informality) according to the *standards of each community*, which are not necessarily assigned by central authorities. This is inherently different than what one witnesses in traditional scholarly publishing and libraries, in which hierarchical systems designate standards and enforce them across their domains of operation.

Turning back to the findings in this study, five participants in this study mentioned they use LaTeX as their primary manuscript preparation tool (in combination with BibTeX), so this indicates the significance of such open document preparation tools to them. However, document preparation does not constitute exploration and discovery but rather enables articulation of research results, so it cannot be considered significant in relation to RQ1.

This leaves us with the chemistry toolkits (KNIME, Open Babel, RDKit, and SciPy) and Xmgrace. The chemistry toolkits were essential and therefore *significant* for Kurt and Gene (I will discuss this directly below). Mary mentioned Xmgrace but did not comment on how she specifically uses it, so I have no data on its *significance* to her research.

What is notable in terms of both the chemistry toolkits and the 2D plotting tool is that they: 1) have been/are created by decentralized peer-production, OSS processes and 2) enable exploration and discovery over *data*. In the case of the chemistry toolkits, this means the exploration of chemical space for drug discovery involving *in-silico* (virtual) testing of molecules which may be biologically active and have "the desired physico-chemical properties to be a drug" (Hoksza and Svozil 2011, p. 201). The key focus here again is data—a scholarly output which does not require authentication in the traditional scholarly publishing or library sense but which rather is informally assigned trust/credibility by the peer community which hosts, curates, or works with it.

In the case of the research being conducted by Kurt and Gene, this kind of exploration and discovery using informal (i.e., peer-created but unauthenticated in the traditional library/archives/scholarly publishing sense) data and open tools is taking place completely beyond the realm of search engines and library infrastructures. Their project is facilitated in the GNAE not by a library or publisher but by MetaCentrum, a national distributed computing infrastructure supported by CESNET, an association of all Czech universities and the Czech Academy of Sciences—not an "institution" in the traditional sense but rather a kind of federation.

Kurt and Gene's research also involves informal communication and collaboration with members of the international open source cheminformatics community, particularly for specialized programming questions regarding open source tools such as RDKit as well as questions about compounds retrieved from public (open) databases. For purposes of comparison between information resources, I removed research groups from the network visualizations in Section 5.2.1.1, but I will reintroduce them here in relation to the following discussion about the *significant allotment of non-institutional/informal information-related exploration and discovery occurring beyond official library-supported mechanisms in the GNAE (RQ1)* for the cheminformaticians in this study. I will highlight notable patterns directly below.

5.2.1.2.1 Informal Exploration and Discovery Pattern One: Importance of Open Data and its Curation

Kurt and Gene, in their project (I will not describe their research in more detail in order not to reveal their identities), originally utilized a set of bioactive compounds which they retrieved from ChEMBL, an open database of over thirteen million “bioactive drug-like small molecules hosted by the European Bioinformatics Institute” which integrates with PubChem and ChemSpider (Van Noorden 2012, p. 524). While such open data enables new forms of drug discovery by more players (i.e., extending their ability to be used beyond pharmaceutical companies with proprietary systems), “the emergence of a number of publicly available bioactivity databases, such as ChEMBL, PubChem, BioAssay and BindingDB, has raised awareness about the topics of data curation, quality and integrity” (Papadatos et al. 2015). In other words, there does not yet exist one overarching system for *assuring authenticity* and the *level of formality* of such data resources available in the GNAE. Despite this, such resources are being used and generally trusted, even with current (and sometimes imperfect) data curation regimes. Distributed peer-community creation and contribution of data to such repositories by scientists themselves (or the organizations they work for) are two reasons for this trust.

Regarding data curation and integrity, Papadatos et al. (2015, n.p.) describe the status of ChEMBL’s processes for these, currently focused on identifying records with possible quality issues as well as on standardizing “activity records, thus making them more accessible and suitable for large-scale data mining and analysis.” The ChEMBL quality assurance process includes both manual and automated components, since “manually checking all publications for cited duplicate values or transcription errors would be an impossible task.” In the future, ChEMBL interestingly might be extended to include include peer community components “in a crowd-sourcing manner, or in-house…according to standardized curation rules” because of the scale of the task (i.e., reviewing millions of chemical structures and related data). An analysis of large scale open data in chemistry would be an interesting area for further study because such data has been openly-available in this discipline for a much shorter time period that in other scientific disciplines and the impact of this has not yet been explored in the scholarly literature, although Long and Schonfeld (2013, pp. 27-31) do discuss smaller collections of chemical research data.

5.2.1.2.2 Informal Exploration and Discovery: Importance of OSS Peers

The opening of data in chemistry has been accompanied by an increase in the number of open source toolkits available to cheminformaticians, including researchers such as Novartis' Greg Landrum, who originally developed such toolkits for commercial application. The important pattern here in relation to *informal* information-related exploration and discovery is that an international community of users of such tools has been created, and community members interact not only at traditional events such as conferences (e.g., RDKitUGM2015, http://www.rdkit.org/) but also *informally* over online discussion forums (mailing lists and Stack Overflow), via monitoring of the websites of prominent researchers/research groups for trends and developments, and in email. The following table provides a summary of these informal peer community resources leveraged for exploration and discovery by Kurt and Gene, with local Czech collaborators removed to protect their privacy. Kurt's commentaries about the nature of these relationships is included, where available.

Table 50: Cheminformatics Community: Informal Peer Resources.

Resource	*Website*	*Type*	*KdS Fieldwork Comments*
Bajorath Research Group	http://www.limes-institut-bonn.de/limes-institut/	Research group	*Publishes a lot of stuff in our field. He visited us in Prague for few days, mutual interest in collaboration.*
Greg Landrum	https://github.com/greglandrum	OSS developer	*Author of RDKit. He is quite active in the mailing lists and KNIME (https://www.knime.org/) community. Writes tutorials and code snippets.*
Laboratoire d'Infochimie	http://infochim.u-strasbg.fr/recherche/group	Research group	*Alexandre Varnek does research on molecular descriptors, pharmacophores, QSAR... the bread and butter of cheminformatics. Also organizes the Strasbourg conference.*
Overington Group	http://www.ebi.ac.uk/research/overington	Research group	*Involved in the ChEMBL database. Loads of useful stuff. [x] knows him personally. He should be in Strasbourg and Hinxton conferences this year.*
Peter Ertl	http://peter-ertl.com/	Prominent Researcher	*He does a lot of cheminformatics in the commercial sphere, our*

			paths have crossed with synthetic feasibility evaluation of arbitrary structures. We had some correspondence regarding the topic, he was very helpful.
Resource	*Website*	*Type*	*KdS Fieldwork Comments*
RDKit discuss	http://sourceforge.net/p/rdkit/mailman/rdkit-discuss/	Discussion forum	n/a
RDKit discuss 2	http://www.mail-archive.com/rdkit-discuss@lists.sourceforge.net/	Discussion forum	n/a
Stack Overflow	http://stackoverflow.com/	Discussion forum	n/a
Tudor Oprea's Flow Cytometry Group	http://hsc.unm.edu/research/flowcyt/index.html	Research group	*Cheminformatics & drug discovery. We haven't attempted to contact him.*
Zinc-fans	http://mailman.bkslab.org/mailman/listinfo		n/a

As was seen in the case of open chemistry data, analyses of the open cheminformatics community do not yet exist in IS literature and this presents an area with future research potential.

5.2.1.3 Summary of RQ1

In the two sections directly above, I presented answers to each component of the primary research question (RQ1) and included some discussion of IS theory in relation to patterns identified thus far (RQ2). Before proceeding into additional theoretical IS and anthropological discussion (RQ2), I can now state that *according to the multi-sited ethnographic analysis of the six scientists in this study, while I observed across all participants a significant allotment of information-related exploration and discovery occurring beyond official library-supported mechanisms in the GNAE, only two participants in the study (both cheminformaticians) illustrated significant non-institutional/informal exploration and discovery activities beyond traditional scholarly publishing/library infrastructures. Such activities included the use of open data and tools to conduct exploration and discovery across open chemistry datasets on a distributed computing infrastructure. Informal activities observed also included communications with a geographically-dispersed community of peers who collaborate on open source tools, share open data, and discuss research topics*.

5.2.1.3.1 Findings Patterns and Prior IS Research

While I have commented on findings of my research in relation to earlier qualitative IS studies above (e.g., Hemminger et al. 2007 and Jamali and Asadi 2010), I would like to briefly highlight several additional patterns from this study's findings in relation to the prior IS studies identified in Section 2.3.4.1. Because my primary research questions involve exploration and discovery beyond library paywalls but not cognitive/information seeking patterns per se, there is a paucity of prior research against which I could compare my results. However, I will at least touch upon a few patterns I observed here.

In this study I found that coding tools have evolved since Allard et al. (2009). Instead of C++ and Java (although Java was mentioned in fieldwork), I observed Kurt and Gene using PostgreSQL and Python, the latter of which Mary also uses. I did not observe a prevalence of internal communication (i.e., use of institutional document repositories or other internal documents) for any participant in this study, but did observe participants using other researchers as sources of information, particularly in the informal interactions identified by Kurt and Gene and in Mary's co-authorship network. A preference for easy access was exhibited across all participants, and quality and trustworthiness were also valued (as in Allard et al. 2009) and illustrated by the overall preference across participants for authenticated, formal resources over informal ones. This coincides with findings by Boukacem-Zeghmouri and Schöpfel (2013, p. 148):

> Some behavior patterns commonly associated with the phenomenon of the Google generation seem rooted in specific disciplinary practice and the culture of science. In fact, some typical so-called Googling behaviors may have pre-Web origins in disciplinary search patterns. In these cases, technology, the Internet and the Web rather accelerated and intensified existing routines than created new information behaviors.

However, the informal data, tools, and communities leveraged by Kurt and Gene illustrate *peer community models of productions*, which are post-Web in origin and which constitute marked changes in information behaviors/practices and concepts of authenticity and trust. Borgman et al. (2012) provide discussion of data sharing and curation by environmental scientists with helpful definitions, but I did not find their description of how environmental scientists and technologists work with data to be comparable to the work conducted by cheminformaticians.

In screenshot data as well as fieldwork, I did observe behaviors as outlined in Ellis et al. (1993), notably: starting (typically with Google or GS), browsing (Google or GS search results supplemented by abstracts), differentiating (selecting article full-text based on abstracts and journals viewed; Dolly's searching session included many differentiation events), monitoring (Kurt mentioned RSS feeds and Dolly, alert services; Mary, Judith, and Sarah would typically use arXiv for this purpose plus specialized journal scanning), verifying (Kurt illustrated this by checking terms for accuracy in Google) and ending (in Dolly's searching session). I did not observe participants in this study chaining (i.e., following citations), except for Kurt's following citations out of Wikipedia into scholarly literature. I found the Ellis model to be most relevant for the purposes of this dissertation because it pragmatically describes the information-seeking activities of individuals and does not involve cognitive aspects and other related issues regarding "information needs" (an analysis of such models is found in Robson and Robinson 2013). I believe such models could be greatly strengthened by future ethnographic research including virtual fieldwork data (video or still images) to assist in testing the theoretical models. Such an analysis is beyond the scope of this dissertation.

For the cheminformaticians, findings in this study generally coincide with Long and Schonfeld (2013) *except* for their conclusion that "chemists have been slow to place their work in online repositories or adopt new publishing models" (p. 5). This was not the case for the two cheminformaticians I observed in this study as well as their community of peers, which has adapted an OSS approach to data, tools, and even their fully-open access (though commercial) *Journal of Cheminformatics*.

Overall, I found prior qualitative IS literature regarding the use of information by scientists to be somewhat outdated, scattered, and difficult to compare the findings of my study against, as did Boukacem-Zeghmouri and Schöpfel (2013, p. 142). I was also unable to locate any prior qualitative IS ethnographic studies which had published associated data and/or codebooks for their projects, and would find that helpful in future research.

5.2.1.3.2 Findings Patterns and Other Social Science Theory

Although I am not an anthropologist by training, I have utilized here multi-sited ethnographic methods. As such, I would like to discuss several social science theories in relation to my findings in order to: 1) counteract the possible "loss of the subaltern" (Section 2.3.4.5) and 2) highlight the importance of tying IS findings to broader theories in anthropology and other social sciences. The research terrain of IS might be greatly enriched if our discipline were less focused on institutional and country boundaries and more open to inquiries about the global implications of changes to search (i.e., Google's dominance in this area) and scholar publishing practices, which Anderson (2015, p. 23) calls "global responsibilities." The scientists I observed in this study are, by necessity, keenly aware the global dimensions of their work; perhaps future IS research will be inspired to move in this direction.

Because I conducted the majority of conventional fieldwork for this study outside the US, at a library situated in the heart of a science and engineering campus, I would like to comment on several patterns I observed in the study related to the dominance (power) of Google, globalization in relation to academic capitalism, the "striving university," and time-space compression.

As Google has begun to dominate the way in which people around the globe search for information, as illustrated in the information-related behaviors of participants in this study, anthropologists and others have begun to investigate what this dominance of one search engine means for human society. Souto-Otero and Beneito-Montagut (2013) describe this kind of power as:

> Being primarily a 'power to' achieve the organizational aims of [search] engine's owner companies rather than a 'power over' others—although elements of this undoubtedly exist. (p. 8)

This indicates a shift in the leadership of information management and an infiltration into the educational sphere: "It is no longer governments, other public institutional stakeholders, or even mass media and publishing houses, but other private corporations (engines' owners) who lead" (p. 9), with a "redrawing of the boundaries between the public and the private, through which private companies penetrate further in educational issues, under the acquiescence of the State, unable or unwilling to regulate global for-profit search engines" (p. 24). Screenshot data gathered in this study do illustrate this kind of penetration of Google into the scholarly environment and its dominance over library alternatives (websites and discovery engines), and traditional libraries have little ability to infiltrate Google's main search interface (or, rather, *interfaces*—which interface one sees depends on many personalization issues combined with local language variants; see Jiang 2014 for a discussion of what this means in countries with national firewalls, such as China).

Although libraries are able to interact with Google Scholar, the level of engagement with each individual library can vary and depends on 1) proper sharing of data with Google by the library, 2) the frequency with which both the library and Google update their systems, and 3) the user activating and properly enabling the GS library settings. As of May 2016, this setting was currently limited to five libraries and is rather hidden on GS's interface, requiring three clicks from the main GS interface plus a proper library search and save for activation.

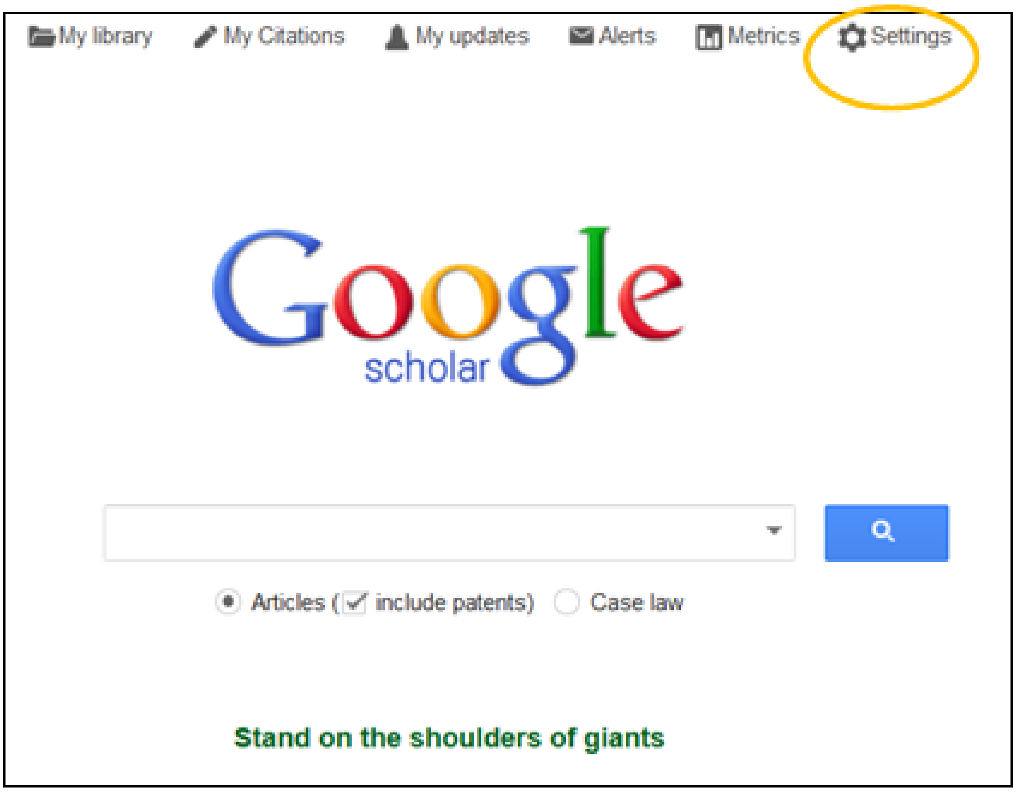


Figure 33: Enabling Library Links: First, One Must Find and Activate Settings

Even if libraries are able to train their users to use these settings properly, they currently have no control over the future existence of this feature; a "Google Library Users' Group" or forum for libraries does not presently exist. And regardless of how much library websites and discovery tools improve in the near-future and how much libraries promote them, it is difficult to imagine that they will be embraced eagerly by scholars, who in this study simply ignored library websites and related search tools. Google is fast, easy, and convenient—I even found myself using it (i.e., the Google CZ variant combined with GS, library links enabled) *more than I ever had before* while writing this dissertation. In contrast, the NTK discovery tool (Summon) seemed increasingly cumbersome the more I used it.

It is difficult to imagine this situation changing in the near-term, which means Google will continue to dominate the way in which scholars find information. The long-term implications of a one search engine world are difficult to contemplate, although there do exist alternatives to Google such as DuckDuckGo and YaCy, the latter of which has no central index but rather is based on a distributed network model (Mager 2014, p. 67). This has led some voices—both those critical of technology developments as well as technology proponents—to call for an independent global public search platform funded internationally or by governments, what Helbing (2014) calls a "planetary nervous system." Regardless of future search scenarios, the only way Google's dominance will change is if a critical number of searchers—and/or non-profit institutions or governments or coalitions of governments or individuals—decide to change this.

One way to do so, as illustrated by the cheminformatics community in this study, is to embrace principles developed by the OSS community, which would indicate the promise of a decentralized open model such as YaCy *if* utilized by a critical number of adherents. These open approaches still baffle economists at both micro- and macro-economic levels and blow existing economic theories apart; Weber (2005) provides an analysis of this. Open access scholarly publishing models, however, do not achieve this effect—they are still based on traditional models of production that, while making content available to broader numbers of people, still often utilize traditional market economy publishing structures rather than the "gift culture" approaches found in well-functioning open source communities (Weber 2005, p. 185).

While an in-depth discussion of globalization is beyond the scope of this dissertation, I do wish to touch upon a few areas which relate to findings in this study.

First, all participants in this study exhibited, in their patterns of information-related exploration and discovery, what Appadurai (2003, p. 45) calls "disjuncture": all six scientists conducted their research in the GNAE concurrently locally and globally within a planetary "circulatory system" of English-language scientific information, with two publications, *Science* and *Nature*, playing a highly prominent roles in the eyes of all participants in my study (and, if I extend this observation to the science and technology campus in which I am embedded, this phenomenon probably ranges much farther), with arXiv trailing—though *prominently* trailing—behind these two science flagships.

Investigating the extent of the power held by *Science* and *Nature* and other English-language databases in scientific information lifeworlds would be an interesting question for future research, but even the findings of this study indicate aspects of globalization and academic capitalism at play. Walker (2009, p. 491) notes:

> Academia is one of the most visible sites of globalization in at least five ways: first, there are large flows of information, ideas, people and courses, and growing numbers of networks formed between people and between institutions; second, it is the home to numerous global cosmopolitan elites who 'know no bounds' (Bauman 2001); third, it can be seen as the birthplace of many new technologies; fourth, it works within and outside the confines of national policy; and lastly, globalization is evident in academia in the numerous cross-country and cross-campus interactions.

Fieldwork and screenshot data in this study did illustrate globalized flows of information and ideas across boundaries facilitated primarily by Google and its regional variants, with disciplinary stars ("elites") published in *Nature* and *Science* and having immediate global impact across disciplines by scientists in different countries who scan their alerts and tables of contents and with scholarly journal distribution occurring over an array of legal gated and open networks as well as cross-national illegal download sites (i.e., LibGen) which are not controlled by governments but are accessed by users in many countries, mostly communicating with one another in English.

I did not observe any significant regional differences between the Americans and Europeans in overall patterns of information-related behavior in this study, except for regional variants of Google's interface and one keyword search in Czech performed by Kurt. All other exploration and discovery I observed took place in global English scientific environments.

For the Czech institutions, this illustrates what Gonzales (2015, p. 303) characterizes as *striving*: "striving" universities (as I observed during conventional fieldwork at the two Czech research universities) attempt to improve international rankings by utilizing the English language and so-called Western models of academic production.

Many of the resources used in this study are provided by institutions or other entities (commercial or non-profit) based in the United States, which indicates its virtual influence for participants in this study. I additionally observed, in relation to online information resources, the phenomenon of time-space compression, in which "globalization is said to seek to annihilate space through time. Space becomes virtual and global transactions occur in 'real time'" (Walker 2009, p. 487). One sees this effect illustrated in the time-space independent interaction of scientists with resources both within and beyond the library paywalls in the GNAE, with resource updates, alerts, and database maintenance typically conducted in global time (i.e., typically in an American time zone which may or may not coincide with the local time patters of resources users). [A]rXiv's founder Paul Ginsparg provides and interesting example of such phenomena, noting how some researchers "time their [arXiv] submissions to arrive just after the daily [US Eastern Standard Time] deadline to maximize their prominence in the next day's mailing" (Ginsparg 2011 and http://arxiv.org/localtime). For someone in Shanghai, this deadline means submitting to arXiv at four in the morning.

Participants in this study did not seem troubled by such trends. The dominance of English-language scholarship and tools was not seen as "colonial" by the Czech and Ukranian native speakers participating in this study, but rather I observed that English acted as an indicator of the *quality* of information (as illustrated in the preference for Wikipedia English) and an indicator of "good science." Kurt and Gene, for example, in several of my conventional fieldwork discussions with them, noted how "local science" which does not get published in international (English language) journals is—in their opinion—often "bad science."

All these topics are ripe for future investigation, in IS and across the social sciences.

5.2.2 RQ3

Taking all of these patterns and theories into consideration, I wanted to exhibit *agency*, which Gonzales (2015, p. 304) citing Archer (2012) calls an act "of resistance, or at least, negotiation intended to advance…notions of the public good against the implications wrought by striving contexts." To do so, I include here a final research question (RQ3): *What might all this mean in the applied sense?* In other words, I will provide examples of how I applied theory and the interim findings of my research to actual services and initiatives created at NTK, the library in which I work, and its surrounding campus during the course of this study in an attempt to close the often-neglected loop between theory, research, and practical application of scholarly findings. The following three examples represent highlights of many other changes implemented in response to my ongoing research.

5.2.2.1 Google and Wikipedia Patterns: Implications for Library Web Design

Because of my pilot work with Mary and early potential participants in the study, I realized early on the challenge library web designers face when competing with Google and Wikipedia on the web. In early 2013, shortly after beginning work at NTK, I was named project leader for the redesign of the NTK website, and the patterns I observed early on in the study—a strong indication of non-library website use, even by mid-career scholars—directly influenced the architecture I selected for the NTK web.

As seen in this study, Google (or GS) search illustrates the importance of utility over form and is characterized by minimal design, rapid and relevant return of results, and very few clicks to full-texts (which are often open, not gated by libraries). GS, for example, has an advantage for serious scholars because of easy import of citation data from search results screens—the user does not have to drill down through a series of library or publisher web pages to find a citation export function.

While the participants in this study started the majority of their searches with Google or GS, NTK's library webpages—according to the library's Google Analytics account in August 2015—still have approximately 5,000 daily accesses on peak days (i.e., semester beginnings and ends). Designing a library website nowadays means serving such users quickly and efficiently and replicating the online experience of online shopping sites and Google. The stakes are high: failure to meet basic user needs means losing them, perhaps forever, beyond the library paywall.

In creating the architecture for NTK's website, my team members and I focused on three main points:

- Minimize clicks to content (i.e., wherever possible, try to make useful content accessible within one click)
- Replicate shopping site functionality, with easy login that indicates to users they have successfully authenticated, even from home. Include online payment options.
- Create parallel design in English—not all pages have to be there, but overall architecture should match that of the Czech pages. This addresses the "striving" nature of the surrounding technology universities, enabling scholars to practice their English, and provides a welcoming environment to the increasing numbers of international students on campus.

A description of the project itself is beyond the scope of this dissertation, but the important point here is that the *redesign architecture was based on early research results with real scientists (i.e., targeted end users)*. The architecture was created as a response to their actual behaviors and needs and was not based on replicating pre-network organizational hierarchies online.

All key points above have now been implemented, and the importance of tying user need to design is apparent. The site launched in March 2014 and, as of August 28, 2015, we have seen a marked increase in both users (+177%) and sessions (+155%), with more page views (+116%) and a decrease in the pages per session (-15%). Regarding the latter, the metric indicates we did not fully achieve our goal of one click per page to substantive content, but we did reduce the number of clicks to content from an average of 2.31 clicks to 1.96 clicks.

The following screenshot highlights features created in response to research observations.

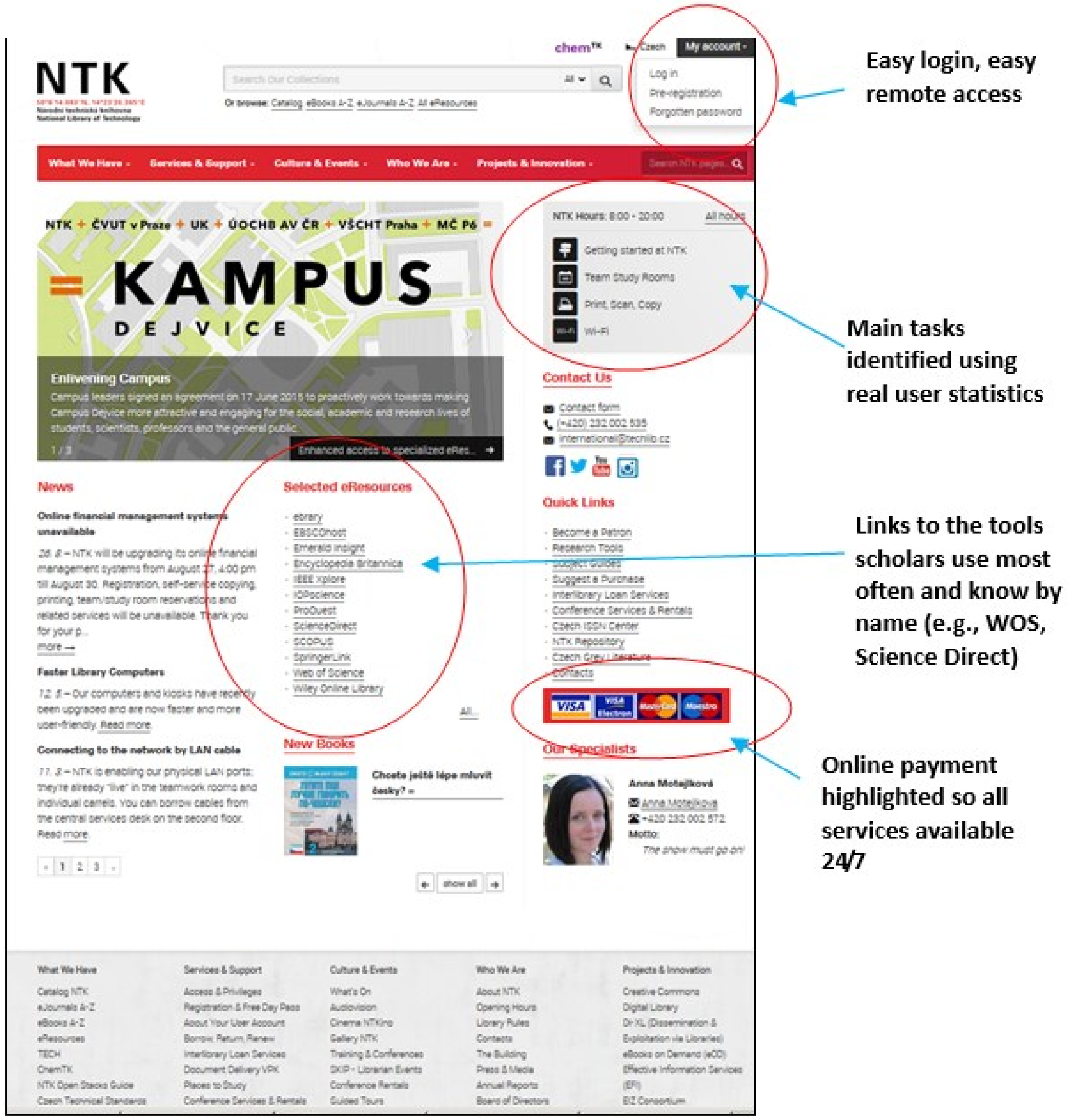


Figure 34: Redesigned techlib.cz

Then, we gave a basic design template to our specialized chemistry colleagues, which they tailored themselves to their own needs (i.e., these *scientists created content for their colleagues*, replicating findings from the OSS community; Figure 35). For the period January 1-May 3, comparing 2015 with 2016, usage for this resource has increased 18% (accesses), with 16% more pageviews averaging two clicks to content (Google Analytics for chemtk.cz, May 3, 2015).

The community for ChemTK is expanding to include the Czech Academy of Sciences Institute for Organic Chemistry and Biochemistry; a screenshot of the next iteration of this site, to be launched in Summer 2016 and even more focused on specific tools and resources, is included below as Figure 36.

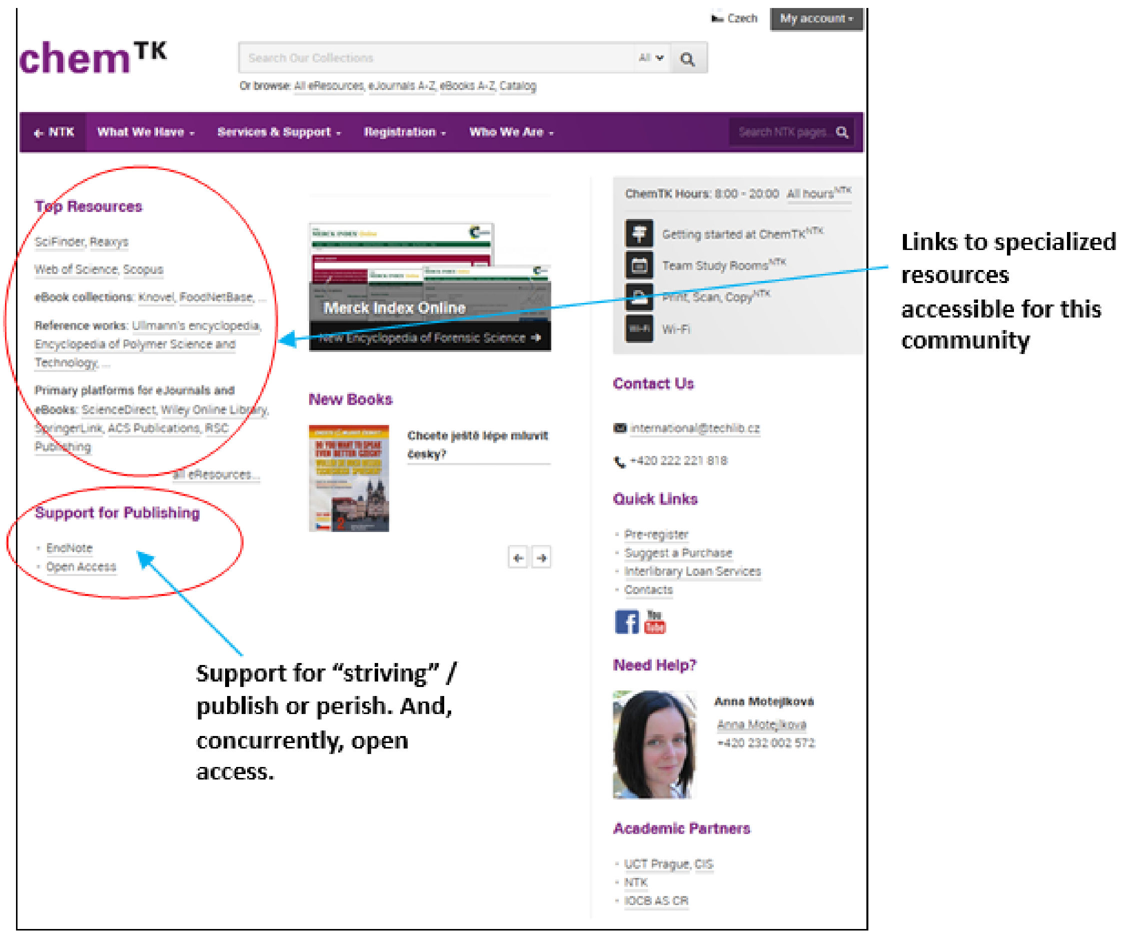


Figure 35: For Scientists by Scientists, chemtk.cz

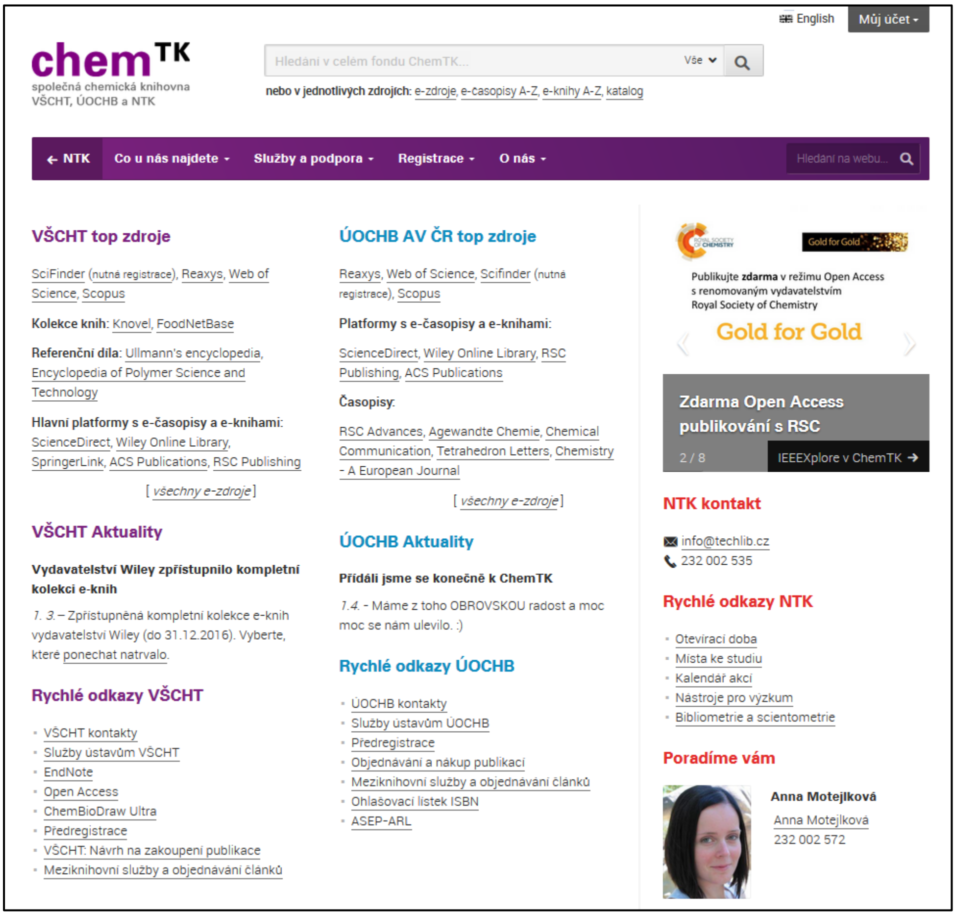


Figure 36: ChemTK, the Next Generation

5.2.2.2 Disintermediation of the Library: Seize Any Instructional Opportunity, However Brief (or: Globalization and Striving: If You Can't Beat Them, Join Them)

As illustrated in this study by Dolly and Mary, enabling library links in GS can dramatically shorten paths to a broad array of full-texts.

Based on such observations as well as observations about scientists' use of document preparation tools such as LaTeX, I created and implemented a four-session pilot course series for doctoral students in civil engineering, working together with a civil engineer and a mathematician. These classes were designed to save professors time in introducing their doctoral students to topics they would otherwise have to teach each doctoral student separately. They were also crafted according to observations in the field about real resource and tool usage patterns and the dominance of English in academic research and publishing. I have included an email with professors regarding topics in Appendix I.

Key topics covered in these English language classes included:

- Remote access to library resources and why they can be more reliable than download sites like LibGen or Sci-Hub
- Activating GS library settings
- Options for working with LaTeX (e.g., including making students aware of JabRef, one of the open source tools Mary mentioned in fieldwork)
- Academic search engine optimization (SEO): Managing your reputation online
- Resources for improving your English
- Publication guidelines for major journals and publishers

The pilot was popular and evaluated well by students and professors, who also attended pilot sessions.

As a result, these professors and I launched a recurring semester-long course on Scientific Writing and Publishing in English open to doctoral students at the Czech Technical University in Prague (CTU) campus in all departments. This course is held in the library but its two instructors are not librarians—one professor specializes in teaching English to Czech native speakers, and the other is a visiting scholar (in 2015-2016, an American Fulbright professor in electrical engineering).

In addition to these classes for doctoral level and higher students, we are now working with several local high schools which have mandatory "bachelor-thesis-like" high school graduation papers (e.g., The Austrian School in Prague). In these classes, we show students not only how to use library resources from home, but how to best utilize Google, GS, Wikipedia, and how to conduct research and write papers in several languages (Czech, English, and German)—all direct responses to the research findings of this study, applicable even for younger learners with non-scientific interests.

5.2.2.3 Open Data, Tools, and OSS Culture: New Staff Competencies, New Ways of Working

As seen in research findings, OSS culture is entering new areas of the GNAE, particularly as data becomes more open and scientists develop new (often open source) tools for working with it, often developed and maintained by decentralized, non-hierarchical peer communities.

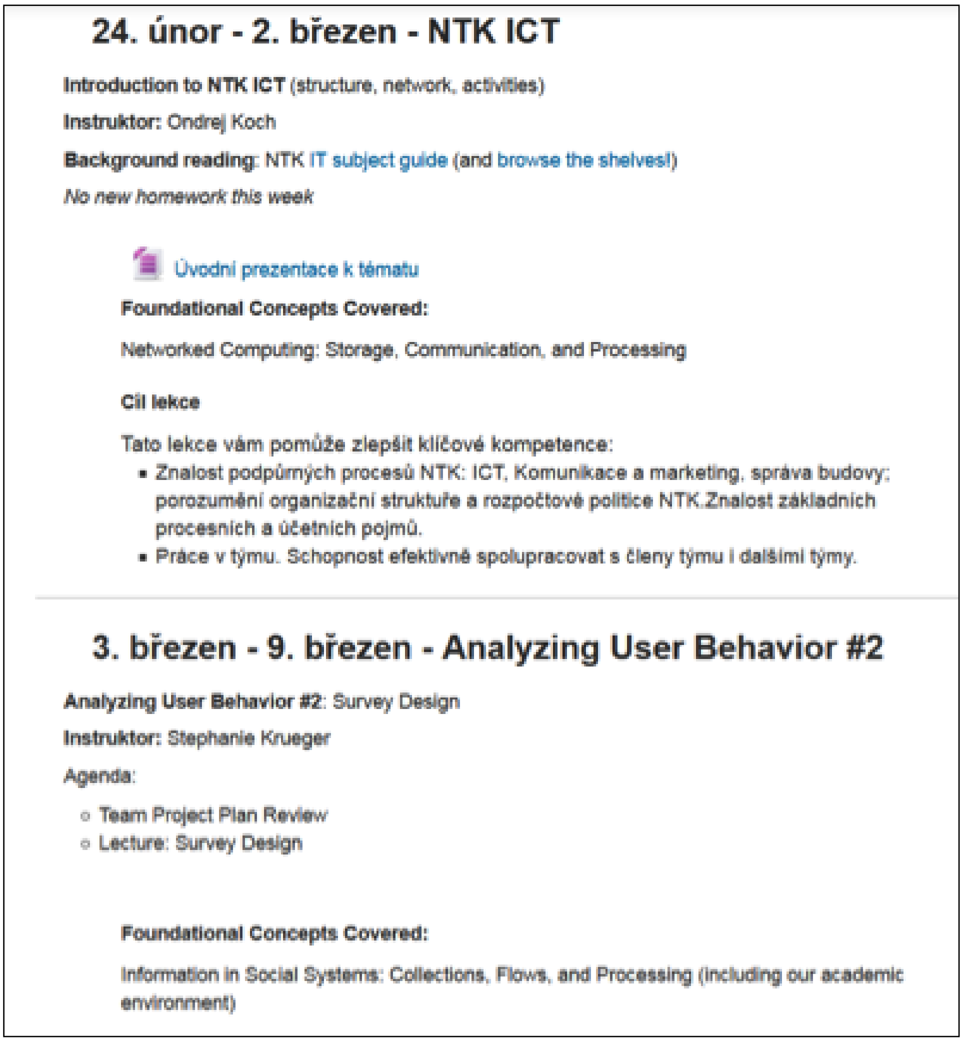
Traditional library cataloging systems and organizational structures conceived in the pre-networked era continue to struggle with changing themselves quickly enough to compete in such an environment. As this study indicates, libraries might appear to many users unfamiliar with them to be *"guardians" of content locked behind a paywall* rather than providers/defenders of broader access to information and knowledge. Sometimes even those working in libraries see themselves in this manner—as *guardians of commercial publisher interests*.

Anderson (2015, p. 21), summarizes the inherent tension of the current situation:

> There is a growing rift between those who believe the library's most fundamental purpose is to support and advance the goals of its host institution and those who believe the library's most important role is as an agent of progress and reform in the larger world of scholarly communication. Although these two areas of endeavor are not mutually exclusive, they are in competition for scarce resources and the choices made between them have serious implications at both the micro level (for the patrons and institutions served by each library) and the macro level (for members of the larger academic community). The tension between these two worldviews is creating friction within librarianship itself: as tightening budgets increasingly force us to choose between worthy programs and projects, there is growing conflict between those whose choices reflect one worldview and those who hold to the other. How this conflict plays out over the next few years may have significant implications for the scholars who depend on libraries for access to research content and for the publishers and other vendors for whom libraries are a core customer base.

At NTK, in order to deal with such issues, we have now implemented a continuous staff education program for selected services team members, to make them aware of the broader environment in which they are working, and to give them the skills to cope with an increasingly complex and technologically-driven environment.

Critical thinking skills play a key role in the curriculum, as do technology, project management, and user-based design skills. In the future, I would like to add analyses of OSS culture and data creation, analysis, and curation techniques to the mix—a direct application of research findings into the class curriculum.



24. únor - 2. březen - NTK ICT

Introduction to NTK ICT (structure, network, activities)

Instruktor: Ondrej Koch

Background reading: NTK IT subject guide (and browse the shelves!)

No new homework this week

Úvodní prezentace k tématu

Foundational Concepts Covered:

Networked Computing: Storage, Communication, and Processing

Cíl lekce

Tato lekce vám pomůže zlepšit klíčové kompetence:

- Znalost podpůrných procesů NTK: ICT, Komunikace a marketing, správa budovy; porozumění organizační struktuře a rozpočtové politice NTK.Znalost základních procesních a účetních pojmů.
- Práce v týmu. Schopnost efektivně spolupracovat s členy týmu i dalšími týmy.

3. březen - 9. březen - Analyzing User Behavior #2

Analyzing User Behavior #2: Survey Design

Instruktor: Stephanie Krueger

Agenda:

- Team Project Plan Review
- Lecture: Survey Design

Foundational Concepts Covered:

Information in Social Systems: Collections, Flows, and Processing (including our academic environment)

Figure 37: Changing Hierarchical Cultures, Sample Internal Training Topics

5.3 Limitations of Study

I have summarized research design and data analysis limitations in Sections 2.3.4.5, 3.1.4, and 4.5.2.4. Regarding overall limitations of this study, because the study included six participants not selected randomly, the findings in this chapter cannot be generalized and found to be representative of broader populations. I have tried to make this clear in the dissertation's narrative, focusing on identifying patterns and areas for future research instead of making broader claims and additionally making comparisons to prior, larger studies whenever possible.

The applied results listed above might be interpreted as some kind of applied generalization, but I intended them rather as examples of how interim research findings can be utilized in applied settings to react more nimbly to the challenges of real work environments. I would make one generalization here: I have found over the course of this project that applying research findings in this way has, to date (May 2016), always yielded positive results.

5.4 Summary of Findings and Discussions

In this chapter, I addressed my primary research questions. I first interpreted data about the information-related behaviors of study participants and observed a general pattern across participants of a significant amount of exploration and discovery taking place beyond library infrastructures. I analyzed this activity, highlighting the apparent importance of Google, Wikipedia English, and open resources and tools. The "amplification" of open resources in the GNAE, for these participants, indicates a disintermediation of the library in exploration and discovery. I then examined patterns of informal/non-institutional exploration and discovery for participants, in the process defining "formal" and "informal" resources, concluding as an answer to primary research questions RQ1 and RQ2 that: *according to the multi-sited ethnographic analysis of the six scientists in this study, while I observed across all participants a significant allotment of information-related exploration and discovery occurring beyond official library-supported mechanisms in the GNAE, only two participants in the study (both cheminformaticians) illustrated significant non-institutional/informal exploration and discovery activities beyond traditional scholarly publishing/library infrastructures. Such activities included the use of open data and tools to conduct exploration and discovery across open chemistry datasets on a distributed computing infrastructure. Informal activities observed also included communications with a geographically-dispersed community of peers who collaborate on open source tools, share open data, and discuss research topics*.

Finally, I provided several examples of how I applied research findings to a real world setting (NTK) in order to answer RQ3, highlighting agile projects designed in relation to interim research findings to library web design, instruction, and internal training.

Chapter 6 Conclusion

In this dissertation, I have provided an example of an adventurous, exploratory research project in IS, utilizing multi-sited ethnographic strategies to address questions about how scientists in different fields really work in rapidly-changing online environments spanning the globe. These information environments exhibit many dual aspects, with local and global, commercial and non-commercial, open and gated, library and non-library information resources all entwined/intertwined and competing on a stage for the attention of scholars.

This study, though not generalizable in the quantitative sense, did identify the following high-level patterns:

- Google's search engine, Google Scholar, and Wikipedia's English versions are important exploration and discovery tools for the scholars observed here, even when English is not their native language. This corresponds to patterns of discovery identified recently in Watkinson et al. (2016, p. 450).
- Libraries, for these participants, do not figure prominently into their discovery and exploration workflows except when access to gated resources is provided seamlessly from an office which does not require remote access. For the two most senior scholars in this study, library remote access mechanisms were completely ignored and avoided, although the scholars were aware of their existence—these scholars appear to conduct the majority of their research *beyond the paywall*.
- Open resources, tools, and data are very important to all participants in this study; none of the participants in this study were skeptical of them, which is the opposite of the phenomonen of *open access suspicion* observed by Watkinson et al. (2016, p. 455).
- Informal exploration and discovery is occurring, notably for the two cheminformaticians, whose research depends on a federated computing infrastructure, open data, and open source software and programming tools. These open resources are often created via peer-community/commons-based models of production, in contrast to traditional scholarly publishing mechanisms established in the pre-network era.

The GNAE, at least for the participants in this study, is dominated by one large commercial player enabling search across them all, Google (i.e., Alphabet Inc.), with rogue players like Sci-Hub and LibGen becoming increasingly popular as alternatives to subscription databases intermediated by libraries (Bohannon 2016). What this means in the long term for scholars, librarians, and publishers is an open question—but the patterns identified above all point towards a continued, accelerating process of disruption and change for all actors involved in the process of scientific exploration and discovery, in which communities of individuals working together across networks around the globe—not only hierarchical institutions—can (and do, in some disciplines such as cheminformatics) play a large and influential role.

Through carefully-designed qualitative research, I believe it is possible to slowly unravel aspects of disruption and change in the GNAE—including deeper understanding of social phenomena such as globalization, academic capitalism, striving universities, and search engine power.

I believe after concluding this initial study that multi-sited ethnographic research strategies involving virtual components are very well-suited to such ambiguous, quickly-shifting information environments, and feel they—particularly when information activities are mapped to network visualizations—can provide valuable insight into phenomena which are often invisible in research projects which do not include comparative components or observations of what is actually happening on people's screens.

I also believe current IS theory, particularly information-seeking theories, could be greatly enriched by studies which address broader contextual issues and which include virtual interaction data.

When appropriate populations can be appropriately defined, I feel it would be helpful to incorporate quantitative analyses into future studies. The global community of cheminformaticians is small enough, for example, that a population could be identified and a random sample of participants generated in order to conduct a larger study similar to this one.

I hope future researchers will be inspired by the prospect of creatively applying research findings from multi-sited ethnographic studies which involve virtual components into real world settings of different kinds in order to develop and launch new services and to explore previously uncharted areas of analysis and debate.

Well-functioning OSS communities do this well: they foster transparent debates in environments where what matters in the end are *applied results* of such dialogues. Weber (2005, p. 186) calls this, discussing OSS culture, applying "human mind space and the commitment of time and intellectual energy by very smart people to a creative enterprise."

What's stopping us from emulating more creative cultures in IS, in libraries—even in global search?

References Cited

AIBAR, E., LLADÓS-MASLLORENS, J., MESEGUER-ARTOLA, A., MINGUILLÓN, J. & LERGA, M. 2015. Wikipedia at University: What Faculty Think and Do About It. *The Electronic Library,* 33**,** 668-683. doi: 10.1108/EL-12-2013-0217

ALLARD, S., LEVINE, K. J. & TENOPIR, C. 2009. Design Engineers and Technical Professionals at Work: Observing Information Usage in the Workplace. *Journal of the American Society for Information Science and Technology,* 60**,** 443-454. doi: 10.1002/asi.21004

ANDERSON, R. 2014. Local and Global, Now and Forever: A Matrix Model of 'Depth Perception' in Library Work. *Against the Grain,* 26**,** 78-79.

ANDERSON, R. 2015. A Quiet Culture War in Research Libraries – and What it Means for Librarians, Researchers and Publishers. *Insights,* 28**,** 21–27. doi: 10.1629/uksg.230

APPADURAI, A. 2003. Illusion of Permanence: Interview with Arjun Appadurai by Perspecta 34. *Perspecta,* 34**,** 44-52. Available: http://www.jstor.org/stable/1567314

ARCHER, M. S. 2012. *The Reflexive Imperative in Late Modernity*, Cambridge, Cambridge University Press.

AMERICAN ANTHROPOLOGY ASSOCIATION. 2012. *Principles of Ethics: Principles of Professional Responsibilities* [Online]. Available: http://users.polisci.wisc.edu/schatzberg/ps919/AAA,%20Ethics,%202012.pdf [Accessed 12/3/2012].

BARABASI, A.-L. (2002). *Linked: The New Science of Networks*. Cambridge, Mass, Perseus Publishing.

BARRY, C. A. 1995. Critical Issues in Evaluating the Impact of IT on Information Activity in Academic Research: Developing a Qualitative Research Solution. *Library and Information Science Research,* 17**,** 107-134. doi: 10.1016/0740-8188(95)90017-9

BAWDEN, D., ROBINSON, L. & SIDDIQUI, T. 2015. “Potentialities or Possibilities”: Towards Quantum Information Science? *Journal of the Association for Information Science and Technology,* 66**,** 437-449. doi: 10.1002/asi.23192

BENKLER, Y. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, New Haven: Yale University Press.

BEYNON-DAVIES, P. 1997. Ethnography and Information Systems Development: Ethnography of, for and within is Development. *Information and Software Technology,* 39**,** 531-540. doi: 10.1016/S0950-5849(97)00008-6

BEYNON-DAVIES, P., MACKAY, H. & TUDHOPE, D. 2000. ‘It’s lots of bits of paper and ticks and post-it notes and things . . .’: A Case Study of a Rapid Application Development Project. *Information Systems Journal,* 10**,** 195-216. Available: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=5466217&lang=cs&site=ehost-live

BLOMBERG, J. & KARASTI, H. 2013. Reflections on 25 Years of Ethnography in CSCW. *Computer Supported Cooperative Work (CSCW),* 22**,** 373-423. doi: 10.1007/s10606-012-9183-1

BODEN, A., MULLER, C. & NETT, B. 2011. Conducting a Business Ethnography in Global Software Development Projects of Small German Enterprises. *Information and Software Technology,* 53**,** 1012-1021. doi: 10.1016/j.infsof.2011.01.009

BOHANNON, J. 2016. Who's Downloading Pirated Papers? Everyone. Science Latest News. Available: http://www.sciencemag.org/news/2016/04/whos-downloading-pirated-

papers-everyone?utm_campaign=news_weekly_2016-04-29&et_rid=140025758&et_cid=453357 [Accessed 5/3/2016].
BORGMAN, C. L., WALLIS, J. C. & MAYERNIK, M. S. 2012. Who's Got the Data? Interdependencies in Science and Technology Collaborations. *Computer Supported Cooperative Work-the Journal of Collaborative Computing,* 21**,** 485-523. doi: 10.1007/s10606-012-9169-z
BÖRNER, K., SANYAL, S. & VESPIGNANI, A. 2007. Network Science. *Annual Review of Information Science and Technology,* 41**,** 537-607. doi: 10.1002/aris.2007.1440410119
BOUKACEM-ZEGHMOURI, C. & SCHÖPFEL, J. 2013. 9 - Beyond the Google Generation: towards Community-specific Usage Patterns of Scientific Information. *In:* BAKER, D. & EVANS, W. (eds.) *Trends, Discovery, and People in the Digital Age.* Whitney (UK), Chandos Publishing.
BOYCE, B. R. & KRAFT, D. H. 1985. Principles and Theories in Information Science. *Annual Review of Information Science and Technology,* 20**,** 153-178.
BRAKKE, K. 2013. *The Surface Evolver* [Online]. Selinsgrove, PA: Ken Brakke. Available: http://www.susqu.edu/brakke/evolver/evolver.html [Accessed 2/28/2013].
BROWN, C. M. 1999. Information Seeking Behavior of Scientists in the Electronic Information Age: Astronomers, Chemists, Mathematicians, and Physicists. *JASIS,* 50**,** 929-943. Available: http://search.proquest.com/docview/231474164?accountid=119841 [Accessed 3/3/2015]
BUCKLAND, M. 2012. What Kind of Science Can Information Science Be? *Journal of the American Society for Information Science and Technology,* 63**,** 1-7. doi: 10.1002/asi.21656
CAIRNS, P. & COX, A. L. 2008. *Research Methods for Human-Computer Interaction,* Cambridge, UK ; New York, Cambridge University Press.
CANDEA, M. 2007. Arbitrary Locations: In Defence of the Bounded Field-Site. *The Journal of the Royal Anthropological Institute,* 13**,** 167-184. doi: 10.1111/j.1467-9655.2007.00419.x
CIBANGU, S. K. 2013. A Memo of Qualitative Research for Information Science: Toward Theory Construction. *Journal of Documentation,* 69**,** 194-213. doi: 10.1108/00220411311300048
CLARK, L., TING, I. H., KIMBLE, C., WRIGHT, P. & KUDENKO, D. 2006. Combining Ethnographic and Clickstream Data to Identify User Web Browsing Strategies. *Information Research-an International Electronic Journal,* 11, 2. Available: http://www.informationr.net/ir/11-2/infres112.html [Accessed 5/1/2014]
CRABTREE, A., NICHOLS, D. M., O'BRIEN, J., ROUNCEFIELD, M. & TWIDALE, M. B. 2000. Ethnomethodologically Informed Ethnography and Information System Design. *Journal of the American Society for Information Science,* 51**,** 666-682. Available: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=3086938&lang=cs&site=ehost-live
CRABTREE, A. E. A. 2012. *Doing Design Ethnography,* London, Springer-Verlag.
CRESWELL, J. W. 2014. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches,* Thousand Oaks, SAGE Publications.
CRONIN, B. 2008. The Sociological Turn in Information Science. *Journal of Information Science,* 34**,** 465-475. doi: 10.1177/0165551508088944

CZECH TECHNICAL UNIVERSITY IN PRAGUE. 2011. *Campus | Study at CTU* [Online]. Prague. Available: http://www.studyatctu.com/campus [Accessed 8/20/2015].

DELEUZE, G. & GUATTARI, F. L. 1987. *A Thousand Plateaus: Capitalism and Schizophrenia,* Minneapolis, University of Minnesota Press.

DICK, B. D. 2010. *The Sociology of Superstring Theory*. PhD thesis. Available: ProQuest Dissertations & Theses Full Text: The Sciences and Engineering Collection. (808293495).

DONG, X., GABRILOVICH, E., HEITZ, G., HORN, W., LAO, N., MURPHY, K., STROHMANN, T., SUN, S. & ZHANG, W. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014. Association for Computing Machinery (ACM), 601-610. Available: https://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf [Accessed 7/2/2015]

DONG, X. L., GABRILOVICH, E., MURPHY, K., DANG, V., HORN, W., LUGARESI, C., SUN, S. & ZHANG, W. 2015. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *Proceedings of the VLDB Endowment,* 8**,** 938-949. Available: http://arxiv.org/pdf/1502.03519.pdf [Accessed 7/2/2015]

EDWARDS, P. N., JACKSON, S. J., CHALMERS, M. K., BOWKER, G. C., BORGMAN, C. L., RIBES, D., BURTON, M. & CALVERT, S. 2013. Knowledge Infrastructures: Intellectual Frameworks and Research Challenges [Online]. Available: http://deepblue.lib.umich.edu/bitstream/2027.42/97552/3/Edwards_etal_2013_Knowledge_Infrastructures.pdf [Accessed 4/16/2014].

ELLIS, D., COX, D. & HALL, K. 1993. A Comparison of the Information Seeking Patterns of Researchers in the Physical and Social Sciences. *Journal of Documentation,* 49**,** 356-369. doi: 10.1108/eb026919

EMARY, L. 2014. When (and When Not) to Use Ethnography for LIS Research. *In:* PRIESTNER, A. A. B., MATT (ed.) *UX in Libraries.* Surrey: Ashgate.

ERNESTO. 2015. *LibGen Goes Down as Legal Pressure Mounts* [Online]. Torrent Freak. Available: https://torrentfreak.com/libgen-goes-down-as-legal-pressure-mounts-150622/ [Accessed 8/25/2015].

ESCOBAR, A., HESS, D., LICHA, I., SIBLEY, W., STRATHERN, M. & SUTZ, J. 1994. Welcome to Cyberia: Notes on the Anthropology of Cyberculture [and Comments and Reply]. *Current Anthropology,* 35**,** 211-231.

FALZON, M.-A. 2012. *Multi-Sited Ethnography: Theory, Praxis and Locality in Contemporary Research*, Farnham, Ashgate Publishing, Ltd.

FIDEL, R. 2008. Are We There Yet?: Mixed Methods Research in Library and Information Science. *Library & Information Science Research,* 30**,** 265-272. doi: 10.1016/j.lisr.2008.04.001

FOUNDATION, C. 2010. *The Carnegie Classification of Institutions of Higher Education™* [Online]. Available: http://carnegieclassifications.iu.edu/ [Accessed 7/30/2015].

FROHMANN, B. Communication Technologies and Human Subjectivity: The Politics of Postmodern Information Science. Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI, 2013. Available: http://www.cais-acsi.ca/ojs/index.php/cais/article/download/569/162 [Accessed 6/4/2014]

FURNER, J. 2004. Information Studies Without Information. *Library Trends,* 52**,** 427-446. Available: https://www.ideals.illinois.edu/bitstream/handle/2142/1684/F?sequence=2

FUTURE, I. F. T. 2001. E-journal Usage and Scholarly Practice. An Ethnographic Perspective on the Role and Impact of e-journal Usage among Users of Biomedical Literature [Online]. Available: http://web.stanford.edu/dept/SUL/library/ejust/findings/full_0801.pdf [Accessed 7/30/2015]
GABRIDGE, T., GASKELL, M. & STOUT, A. 2008. Information Seeking through Students' Eyes: The MIT Photo Diary Study. *College & Research Libraries,* 69**,** 510-523. doi: 10.5860/crl.69.6.510
GARCIA, A. C., STANDLEE, A. I., BECHKOFF, J. & CUI, Y. 2009. Ethnographic Approaches to the Internet and Computer-Mediated Communication. *Journal of Contemporary Ethnography,* 38**,** 52-84. Available: http://www.grupciber.net/blog/wp-content/uploads/2012/11/6-Ethnographic-Approaches-to-the-Internet-and-Computer-Mediated-Communication.pdf
GEERTZ, C. 1987. The 'Interpretation of Cultures' - Thick Description toward an Interpretative Theory of Culture. *Aut Aut***,** 151-176. Available: http://people.ucsc.edu/~ktellez/geertz1973.pdf [Accessed 11/3/2012]
GINSPARG, P. 2011. ArXiv at 20. *Nature,* 476**,** 145-147. doi: 10.1038/476145a
GLÄSER, J. 2006. *Wissenschaftliche Produktionsgemeinschaften - die soziale Ordnung der Forschung,* Frankfurt/New York, Campus Verlag.
GONZALES, L. D. 2015. Faculty Agency in Striving University Contexts: Mundane Yet Powerful Acts of Agency. *British Educational Research Journal,* 41**,** 303-323. doi: 10.1002/berj.3140
GOODMAN, L. A. 1961. Snowball Sampling. *The Annals of Mathematical Statistics***,** 148-170. Available: http://projecteuclid.org/euclid.aoms/1177705148 [Accessed 5/4/2014]
GOODMAN, L. A. 2011. Comment: On Respondent-Driven Sampling and Snowball Sampling In Hard-To-Reach Populations And Snowball Sampling Not In Hard-To-Reach Populations. *Sociological Methodology,* 41**,** 347-353. doi: 10.1111/j.1467-9531.2011.01242.x
HAAS, C. 2014. Wozu Ethnografie in Bibliotheken? *Bibliothek Forschung und Praxis,* 38**,** 185-189. Available: http://www.degruyter.com/dg/viewarticle/j$002fbfup.2014.38.issue-2$002fbfp-2014-0023$002fbfp-2014-0023.xml [Accessed 6/25/2015]
HARTEL, J. & THOMSON, L. 2011. Visual Approaches and Photography for the Study of Immediate Information Space. *Journal of the American Society for Information Science and Technology,* 62**,** 2214-2224. doi: 10.1002/asi.21618
HEATH, C. & HINDMARSH, J. 2002. Analysing Interaction. Video, Ethnography and Situated Conduct. *In:* MAY, T. (ed.) *Qualitative Research in Practice*. London, Sage.
HELBING, D. 2014. The World after Big Data: What the Digital Revolution Means for Us [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2438957 [Accessed 7/30/2015].
HEMMINGER, B. M., LU, D., VAUGHAN, K. & ADAMS, S. J. 2007. Information Seeking Behavior of Academic Scientists. *Journal of the American Society for Information Science and Technology,* 58**,** 2205-2225. doi: 10.1002/asi.20686
HINE, C. 2000. *Virtual Ethnography*, Newcastle upon Tyne, Sage.
HINE, C. 2007. Multi-Sited Ethnography as a Middle Range Methodology for Contemporary STS. *Science, Technology, & Human Values,* 32**,** 652-671. doi: 10.1177/0162243907303598

HINE, C., ED. 2006. *New Infrastructures for Knowledge Production: Understanding e-Science,* Hershey, Information Science Publishing.
HOKSZA, D. & SVOZIL, D. Exploration of Chemical Space by Molecular Morphing. Bioinformatics and Bioengineering (BIBE), 2011 IEEE 11th International Conference on, 2011. IEEE, 201-208. doi: 10.1109/BIBE.2011.38
HOKSZA, D., ŠKODA, P., VORŠILÁK, M. & SVOZIL, D. 2014. Molpher: A Software Framework for Systematic Chemical Space Exploration. *Journal of Cheminformatics,* 3**,** 32. doi: 10.1186/1758-2946-6-7
HOWARD, P. N. 2002. Network Ethnography and the Hypermedia Organization: New Media, New Organizations, New Methods. *New Media & Society,* 4**,** 550-574. doi: 10.1177/146144402321466813
HUBER, G. L. & GÜRTLER, L. 2014. *AQUAD* [Online]. Available: http://www.aquad.de/ [Accessed August 12, 2014].
JAMALI, H. R. & ASADI, S. 2010. Google and the Scholar: The Role of Google in Scientists' Information-seeking Behaviour. *Online Information Review,* 34**,** 282-294. doi: 10.1108/14684521011036990
JIANG, M. 2014. The Business and Politics of Search Engines: A Comparative Study of Baidu and Google's Search Results of Internet Events in China. *New Media & Society,* 16**,** 212-233. doi: 10.1177/1461444813481196
JIROTKA, M., LEE, C. P. & OLSON, G. M. 2013. Supporting Scientific Collaboration: Methods, Tools and Concepts. *Computer Supported Cooperative Work (CSCW),* 22**,** 667-715. doi: 10.1007/s10606-012-9184-0
KHOO, M., ROZAKLIS, L. & HALL, C. 2012. A Survey of the Use of Ethnographic Methods in the Study of Libraries and Library Users. *Library & Information Science Research,* 34**,** 82-91. doi: 10.1016/j.lisr.2011.07.010
KJELDSKOV, J. & STAGE, J. 2012. Combining Ethnography and Object-Orientation for Mobile Interaction Design: Contextual Richness and Abstract Models. *International Journal of Human-Computer Studies,* 70**,** 197-217. doi: 10.1016/j.ijhcs.2011.10.004
LATOUR, B., WOOLGAR, S. & SALK, J. 1986. *Laboratory Life: The Construction of Scientific Facts,* Princeton, Princeton University Press.
LEVINE, K. J., ALLARD, S. & TENOPIR, C. 2011. The Changing Communication Patterns of Engineers. *Proceedings of the IEEE,* 99**,** 1155-1157. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5888502 [Accessed 4/12/2015]
LEVITT, S. D. & LIST, J. A. 2011. Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. *American Economic Journal: Applied Economics,* 3**,** 224-238. doi: 10.1257/app.3.1.224
LINGEL, J. 2015. Information Practices of Urban Newcomers: An Analysis of Habits and Wandering. *Journal of the Association for Information Science and Technology,* 66**,** 1239-1251. doi: 10.1002/asi.23255
LONG, M. P. & SCHONFELD, R. C. 2013. Supporting the Changing Research Practices of Chemists [Online]. Available: http://www.sr.ithaka.org/wp-content/uploads/2015/08/Supporting-the-Changing-Research-Practices-of-Chemists-FINAL.pdf [Accessed 12/16/2013].
LUFT, J. 2015. The Challenges of Being a Fox – Library and Information Science as an Applied Discipline. *Bibliothek Forschung und Praxis,* 39**,** 132-137. Available:

http://www.degruyter.com/view/j/bfup.2015.39.issue-2/bfp-2015-0015/bfp-2015-0015.xml [Accessed 6/25/2015]

LUO, Y. 2014. *Use of Web 2.0 Technologies: A Virtual Ethnographic and Phenomenological Study of First-Year Engineering Students' Experiences.* PhD thesis, Purdue University. Available: http://docs.lib.purdue.edu/dissertations/AAI3636377/ [Accessed 2/27/2015]

MARCUS, G. E. 1995. Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography. *Annual Review of Anthropology,* 24**,** 95-117. Available: http://www.jstor.org/stable/2155931

MARTIN, M. 1993. Geertz and the Interpretive Approach in Anthropology. *Synthese,* 97**,** 269-286. doi: 10.1007/BF01064118

MATTHEWS, R. J. & BAISH, M. A. 2007. State-By-State Report on Authentication of Online Legal Resources. Washington, DC: American Association of Law Libraries [Online]. Available: http://www.aallnet.org/mm/Publications/products/gm-node/24485.aspx [Accessed 3/12/2015].

MCEWEN, R. S., KATHLEEN 2012. Orality in the Library: How Mobile Phones Challenge our Understandings of Collaboration in Hybridized Information Centers. *Library and Information Science Research,* 34**,** 92-98. Available: http://hdl.handle.net/2142/15021

MILLS, D. & RATCLIFFE, R. 2012. After Method? Ethnography in the Knowledge Economy. *Qualitative Research,* 12**,** 147-164. doi: 10.1177/1468794111420902

NADAI, E. & MAEDER, C. 2005. Fuzzy Fields. Multi-Sited Ethnography in Sociological Research. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research,* 6. Available: http://www.qualitative-research.net/index.php/fqs/article/viewArticle/22

NTK. 2015. *About NTK: History* [Online]. Prague: Czech National Library of Technology. Available: https://www.techlib.cz/en/2749-about-ntk#tab_history [Accessed 7/31/ 2015].

ORTNER, S. B. 2007. Clifford Geertz (1926-2006). *American Anthropologist,* 109**,** 786-789. doi: 10.1525/aa.2007.109.4.786

PAPADATOS, G., GAULTON, A., HERSEY, A. & OVERINGTON, J. P. 2015. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. *Journal of Computer-aided Molecular Design***,** 1-12. doi: 10.1007/s10822-015-9860-5

PEPE, A. 2010. *Structure and Evolution of Scientific Collaboration Networks in a Modern Research Collaboratory.* PhD thesis, University of California. doi: 10.2139/ssrn.1616935

PEPE, A. 2011. The Relationship Between Acquaintanceship and Coauthorship in Scientific Collaboration Networks. *Journal of the American Society for Information Science and Technology,* 62**,** 2121-2132. doi: 10.1002/asi.21629

PEPE, A. & MAYERNEK, M. 2012. The Use of Micro-Blogging for Field-based Scientific Research. *Proceedings of the 45th Hawaii International Conference on System Science (HICSS-45 2012).* Maui, Hawaii. doi: 10.1109/HICSS.2012.589

PRIOR, D. D. & MILLER, L. M. 2012. Webethnography Towards a Typology for Quality in Research Design. *International Journal of Market Research,* 54**,** 503-520. doi: 10.2501/IJMR-54-4-503-520

RHINESMITH, C. H. 2014. *The Social Shaping of Cloud Computing: An Ethnography of Infrastructure in East St. Louis, Illinois.* PhD thesis, University of Illinois at Urbana-

Champaign. Available: http://search.proquest.com/pqdtscieng/docview/1651961259 [Accessed 11/3/2013]
RIEMER, F. 2008. Addressing Ethnographic Inquiry. *Research Essentials.* San Francisco, Jossey-Bass Publishers.
ROBBINS, S. E., DEBRA; KULP, CHRISTINA 2011. How Unique Are Our Users? Comparing Responses Regarding the Information-Seeking Habits of Engineering Faculty. *College & Research Libraries,* 72**,** 515-532. doi: 10.5860/crl-156
ROBSON, A. & ROBINSON, L. 2013. Building on Models of Information Behaviour: Linking Information Seeking and Communication. *Journal of Documentation,* 69**,** 169-193.
ROSENBLUM, E. L. 2015. Librarians are Already in the Field: How and Why to Begin Ethnographic Fieldwork. *Bibliothek Forschung und Praxis,* 39**,** 38-142. Available: http://www.degruyter.com/view/j/bfup.2015.39.issue-2/bfp-2015-0016/bfp-2015-0016.xml
ROWLANDS, I., NICHOLAS, D., WILLIAMS, P., HUNTINGTON, P., FIELDHOUSE, M., GUNTER, B., WITHEY, R., JAMALI, H. R., DOBROWOLSKI, T. & TENOPIR, C. The Google Generation: The Information Behaviour of the Researcher of the Future. Aslib Proceedings, 2008. Emerald Group Publishing Limited, 290-310. doi: 10.1108/00012530810887953
RUSS, A.R. 2012. *The Illustion of History: Time and the Radical Political Imagination*. Washington, DC, The Catholic University of American Press.
SANDSTORM, A. R. & SANDSTROM, P. E. 1995. The Use and Misuse of Anthropological Methods in Library and Information Science Research. *The Library Quarterly: Information, Community, Policy,* 65**,** 161-199. Available: http://www.jstor.org/stable/4309020
SARACEVIC, T. 2010. Information Science. *In:* BATES, M. J. M., MARY NILES (ed.) *Encyclopedia of Library and Information Sciences, Third Edition.* Third ed. Boca Raton: Taylor & Francis.
SEADLE, M. 2011. Research Rules for Library Ethnography. *Library Hi Tech,* 29**,** 409-411. doi: 10.1108/07378831111174378
SEELIGER, F. 2013. „Individuum est ineffabile "oder: Verstehen Bibliothekare ihre Zielgruppe wirklich? *027.7 Zeitschrift für Bibliothekskultur/Journal for Library Culture,* 1. doi: 10.12685/027.7-1-3-36
SIMONITE, T. 2013. *The Decline of Wikipedia* [Online]. Cambridge, MA: MIT Technology Review. Available: http://www.technologyreview.com/featuredstory/520446/the-decline-of-wikipedia/ [Accessed August 24, 2015].
SIMPSON, T. W. 2012. Evaluating Google as an Epistemic Tool. *Metaphilosophy,* 43**,** 426-445. doi: 10.1111/j.1467-9973.2012.01759.x
SKÅGEBY, J. 2012. The Irony of Serendipity: Disruptions in Social Information Behaviour. *Library Hi Tech,* 30**,** 321-334. doi: 10.1108/07378831211239988
SKENDERIJA, S. The Media Paradigm for Information Science. Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI, 1999. Available: http://scholarship.law.cornell.edu/clsops_papers/37
SKENDERIJA, S. 2008. Miroslav Petricek and the Quest for a New Ontology of Information. *Cornell Law Faculty Working Papers. Paper 34.* Available: http://scholarship.law.cornell.edu/clsops_papers/34
SKENDERIJA, S. 2009. Permanent Digital Legal Information: Disappearing Urls and Preservation of Digital Objects Cited in Court Decisions. *International Journal of Legal*

Information, 37**,** 315-320. Available: http://scholarship.law.cornell.edu/ijli/vol37/iss3/6
SOUTO-OTERO, M. & BENEITO-MONTAGUT, R. 2013. 'Power on': Googlecracy, Privatisation and the Standardisation of Sources. *Journal of Education Policy,* 28**,** 481-500. doi: 10.1080/02680939.2013.775347
SPINNEY, J. 2011. A Chance to Catch a Breath: Using Mobile Video Ethnography in Cycling Research. *Mobilities,* 6**,** 161-182. doi: 10.1080/17450101.2011.552771
STODDEN, V. 2009. Enabling Reproducible Research: Open Licensing for Scientific Innovation. *International Journal of Communications Law and Policy* [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1362040 [Accessed 4/12/2015].
TECHSMITH. 2015. *Jing, Free Screenshot and Screencast Software* [Online]. Available: https://www.techsmith.com/jing.html [Accessed 1/5/2014].
TENOPIR, C. & KING, D. W. 2004. *Communication Patterns of Engineers,* Hoboken, NJ, John Wiley.
TEPLITSKIY, M., LU, G. & DUEDE, E. 2015. Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science. Available: arXiv: 1506.07608
TSEKERIS, C. 2008. Thoughts on the Nature of the Virtual. *Ubiquity,* 2008. doi: 10.1145/1399625.1399626
VAN NOORDEN, R. 2012. Chemistry's Web of Data Expands. *Nature,* 483**,** 524. Available: https://www.openphacts.org/documents/publications/Van%20Noorden_Chemistrys%20web%20of%20data%20expands_mentioning%20OPS_Nature%20Vol%20483_p%20524_120329.pdf
VELDEN, T. & LAGOZE, C. 2013. The Extraction of Community Structures from Publication Networks to Support Ethnographic Observations of Field Differences in Scientific Communication. *Journal of the American Society for Information Science and Technology,* 64**,** 2405-2427. doi: 10.1002/asi.22929
WALKER, J. 2009. Time as the Fourth Dimension in the Globalization of Higher Education. *The Journal of Higher Education,* 80**,** 483-509. doi: 10.1353/jhe.0.0061
WALSH, J. 2014. *The Highlight Reel and the Real Me: How Adolescents Construct the Facebook Fable.* PhD thesis, Boston University. Available: http://hdl.handle.net/2144/15407
WATKINSON, A., NICHOLAS, D., THORNLEY, C., HERMAN, E., JAMALI, H. R., VOLENTINE, R., ALLARD, S., LEVINE, K., TENOPIR, C. (2016). Changes in the Digital Scholarly Environment and Issues of Trust: An Exploratory, Qualitative Analysis. *Information Processing & Management*, 52, 3, 446-458. doi: 10.1016/j.ipm.2015.10.002
WEBER, S. 2005. The Political Economy of Open Source Software and Why it Matters. *Digital Formations: IT and New Architectures in the Global Realm***,** 178-213. Princeton, Princeton University Press.
WEST, R., WEBER, I. & CASTILLO, C. Drawing a Data-Driven Portrait of Wikipedia Editors. Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, 2012. Association for Computing Machinery (ACM), 3. doi: 10.1145/2462932.2462937
WHITTLE, A. 2000. Ethnography on the Move: From Field to Net to Internet. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research.* Available: http://www.qualitative-research.net/index.php/fqs/article/viewArticle/1131

YE, F. Y., YU, S. S. & LEYDESDORFF, L. 2013. The Triple Helix of University-Industry-Government Relations at the Country Level and its Dynamic Evolution under the Pressures of Globalization. *Journal of the American Society for Information Science and Technology,* 64**,** 2317-2325. doi: 10.1002/asi.22931
YOON, A. 2014. End Users' Trust in Data Repositories: Definition and Influences on Trust Development. *Archival Science,* 14**,** 17-34. doi: 10.1007/s10502-013-9207-8
YOSHIKAWA, H., KALIL, A., WEISNER, T. S. & WAY, N. 2008. Mixing Qualitative and Quantitative Research in Developmental Science: Uses and Methodological Choices. *Developmental Psychology,* 44**,** 344-354. doi: 10.1037/0012-1649.44.2.344
ZHOU, X. 2010. *Information in Healthcare: An Ethnographic Analysis of a Hospital Ward.* PhD thesis, University of Michigan. Available: https://deepblue.lib.umich.edu/handle/2027.42/78940

Appendix A: Illustration of Very Few Marcus Citations in Information Science

Web of Science Citations of Marcus (1995) in INFORMATION SCIENCE LIBRARY SCIENCE category (data as retrieved/send from WOS)

AU Pollock, N
  Williams, R
AF Pollock, Neil
  Williams, Robin
TI The venues of high tech prediction: Presenting the future at industry
  analyst conferences
SO INFORMATION AND ORGANIZATION
AB This paper attempts to understand the apparent paradox that although industry analyst information technology (IT) predictions often turn out to be 'wrong', there appears no obvious decline in the number of predictions made, the appetite for this kind of knowledge, or the standing of those producing this kind of insight. This begs the following questions: How do industry analysts come up with predictions? Who or what is involved in their shaping? How do they establish their efficacy? How do they and others evaluate these predictions? And what value do they have for those who consume them? We have been able to examine these issues empirically through ethnographic study of one of the key interfaces between the production and consumption of predictions: the industry analyst conference. In departing from studies that foreground its 'accuracy', we describe how this knowledge is subject to more plural methods of evaluation and accountability concerning its utility. We show how industry analysts gauge the utility of their knowledge through interacting with and provoking reactions from conference audiences. We analyse these interactions not simply as a means to socialise this knowledge but as a space for the simultaneous production and validation of predictions and the role of the audience as offering a new form of 'public proof'. We also describe how these conferences have led to a reshaping of the kinds of experts and expertise involved in producing and communicating this knowledge. Our material is based on interviews with a number of industry analysts and observations of the conferences of the leading industry analyst firm Gartner Inc. (C) 2015 Elsevier Ltd. All rights reserved.
SN 1471-7727

DI 10.1016/j.infoandorg.2015.04.001
UT WOS:000356105400003

AU Cresswell, KM
  Bates, DW
  Williams, R
  Morrison, Z
  Slee, A
  Coleman, J
  Robertson, A
  Sheikh, A
AF Cresswell, Kathrin M.

Bates, David W.
Williams, Robin
Morrison, Zoe
Slee, Ann
Coleman, Jamie
Robertson, Ann
Sheikh, Aziz
TI Evaluation of medium-term consequences of implementing commercial computerized physician order entry and clinical decision support prescribing systems in two 'early adopter' hospitals
SO JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION
AB Objective To understand the medium-term consequences of implementing commercially procured computerized physician order entry (CPOE) and clinical decision support (CDS) systems in 'early adopter' hospitals.

Materials and methods In-depth, qualitative case study in two hospitals using a CPOE or a CDS system for at least 2 years. Both hospitals had implemented commercially available systems. Hospital A had implemented a CPOE system (with basic decision support), whereas hospital B invested additional resources in a CDS system that facilitated order entry but which was integrated with electronic health records and offered more advanced CDS. We used a combination of documentary analysis of the implementation plans, audiorecorded semistructured interviews with system users, and observations of strategic meetings and systems usage.

Results We collected 11 documents, conducted 43 interviews, and conducted a total of 21.5 h of observations. We identified three major themes: (1) impacts on individual users, including greater legibility of prescriptions, but also some accounts of increased workloads; (2) the introduction of perceived new safety risks related to accessibility and usability of hardware and software, with users expressing concerns that some problems such as duplicate prescribing were more likely to occur; and (3) realizing organizational benefits through secondary uses of data.

Conclusions We identified little difference in the medium-term consequences of a CPOE and a CDS system. It is important that future studies investigate the medium- and longer-term consequences of CPOE and CDS systems in a wider range of hospitals.
SN 1067-5027
EI 1527-974X
PD OCT
PY 2014
VL 21
IS E2
BP 194
EP 202
DI 10.1136/amiajnl-2013-002252

AU Jaclin, D
AF Jaclin, David
TI In the (bleary) eye of the tiger: An anthropological journey into jungle backyards

SO SOCIAL SCIENCE INFORMATION SUR LES SCIENCES SOCIALES
AB North America shelters a growing population of so-called 'exotic animals'. If the phenomenon is not recent, it now fuels a considerable black market. Jungle backyards compose a non-negligible (yet often neglected) part of some modern ecological landscapes. This article explores problematical situations emerging from these shared humanimal lives. It presents the first results of a multi-species ethnography and examines the prevalence of what I call beastness - an antique commerce amid humans and animals that reveals not only utilitarian purposes, but also relational entanglements. Such a commerce feeds a sizeable economy and exerts major selective pressures (both biological and cultural) on organisms and their environment. For instance, there are more captive tigers living in the state of Texas alone than wild specimens running free anywhere else on the planet. From a strictly statistical point of view, the average tiger is no longer the tiger we imagine. Not wild anymore but neither quite domesticated, some animals - pioneers, in a sense - shuffle traditional taxonomical and ontological conceptions. Through biographical material, I reflect on adaptive responses as well as on zoological potentialities developed by this always-evolving bestiary. Providing serious case studies to further debates dealing with bio-eco-conservation, I discuss the influence of informational and communicational processes crystallized by some of our contemporary crossed becomings.
SN 0539-0184
EI 1461-7412
PD JUN
PY 2013
VL 52
IS 2
BP 257
EP 271
DI 10.1177/0539018413477522

AU Williams, R
Pollock, N
AF Williams, Robin
Pollock, Neil
TI Moving Beyond the Single Site Implementation Study: How (and Why) We Should Study the Biography of Packaged Enterprise Solutions
SO INFORMATION SYSTEMS RESEARCH
AB The single site implementation study is an invaluable tool for studying the large-scale enterprise solution. Together with constructivist frameworks and ethnographic approaches it has allowed the development of rich local pictures of the immediate and adaptive response by user organizations to the take-up of what are, today, often generic packaged systems. However, to view the packaged enterprise solution principally at the place where the user encounters it also has limitations. It produces somewhat partial understandings of these complex artifacts. In particular, it downplays important influences from other sites and time frames. This paper argues that if we are to understand the full implications of enterprise solutions for organizations then we should study their "biography." This idea points to how the

career of workplace technology is often played out over multiple time frames and settings. To understand its shaping therefore requires scholars to go beyond the study of technology at a single locale or moment and, rather, attempt to follow it through space and time. The paper develops two ideas to aid this kind of study. We discuss better spatial metaphors that might help us explore the hybrid and extended spaces in which packaged software systems develop and evolve. We also review improved temporal understandings that may capture the multiple contemporary and historical time frames at play. The paper concludes by discussing some possible research directions that a focus on the biography of a technology might allow.
SN 1047-7047
PD MAR
PY 2012
VL 23
IS 1
BP 1
EP 22
DI 10.1287/isre.1110.0352

AU Palen, L
   Vieweg, S
   Liu, SB
   Hughes, AL
AF Palen, Leysia
   Vieweg, Sarah
   Liu, Sophia B.
   Hughes, Amanda Lee
TI Crisis in a Networked World Features of Computer-Mediated Communication
   in the April 16, 2007, Virginia Tech Event
SO SOCIAL SCIENCE COMPUTER REVIEW
AB Crises and disasters have micro and macro social arrangements that differ from routine situations, as the field of disaster studies has described over its 100-year history. With increasingly pervasive information and communications technology and a changing political arena where terrorism is perceived as a major threat, the attention to crisis is high. Some of these new features of social life have created changes in disaster response that we are only beginning to understand. The University of Colorado is establishing an area of sociologically informed research and information and communications technology development in crisis informatics. This article reports on research that examines features of computer-mediated communication and information sharing activity during and after the April 16, 2007, crisis at Virginia Tech by members of the public. The authors consider consequences that these technology-supported social interactions have on emergency response and implications for methods in e-Social Science.
SN 0894-4393
PD NOV
PY 2009
VL 27
IS 4

BP 467
EP 480
DI 10.1177/0894439309332302

PT J
AU Rye, SA
AF Rye, Stale Angen
TI Negotiating the Symbolic Power of Information and Communication
  Technologies (ICT): The Spread of Internet-Supported Distance Education
SO INFORMATION TECHNOLOGY FOR DEVELOPMENT
AB The Internet may be, as typically suggested, important in distance education for facilitating connections between groups of students, educational institutions, and external learning resources. This article, however, reveals why this is not the only reason for applying information and communication technologies (ICT) in higher education in a remote area in a developing country. In addition, the Internet seems to be of great importance in symbolizing modernization and progress, thereby adding symbolic power to such education. Empirical sources originate from an explorative case study of an Internet-supported distance education program in the province of Bangka Belitung in Indonesia. Based on a translation perspective on the spread of pheromones, the analyses of empirical sources show how the Internet has contributed to the spread of distance education, but paradoxically this has not had much effect on the use of Internet by students in peripheral areas, at least not in the short term. (C) 2008 Wiley Periodicals, Inc.
SN 0268-1102
PY 2009
VL 15
IS 1
BP 17
EP 31
DI 10.1002/itdj.20110

AU Ostrander, M
AF Ostrander, Margaret
TI Talking, looking, flying, searching: information seeking behaviour in
  Second Life
SO LIBRARY HI TECH
AB Purpose - This research seeks to answer, "How do everyday Second Life users go about finding needed information?" as the primary research question.
  Design/methodology/approach - A virtual ethnographic approach couched in grounded theory was utilized to conduct semi-structured interviews with everyday users of Second Life, accompanied by participant observation.
  Findings - Information seeking behaviors within the virtual world of Second Life were found to be rich, complex interaction with multiple facets. Five themes emerged to illuminate how users seek information.
  Research limitations/implications - This research took place over a six-week period, although most enthographies last at least one year. Conclusions were drawn solely from interviews because participant observation did not penetrate a given

community with enough depth to adequately address the research question.
Practical implications - Virtual worlds offer the promise of becoming an integrated part of the information seeking landscape for an increasing number of users. Understanding the factors influencing information seeking behavior that are outlined in this article will equip librarians and information professionals to best utilize virtual worlds and continue to create innovative, user-focused services there.
Originality/value - This article extends current scholarship by offering a practical, five-factor approach to understand how people seek information in virtual worlds. The literature is robust in description about library services and the nature of information in virtual worlds. Yet, investigation into information seeking behavior in this environment is in its nascent stages.
SN 0737-8831
PY 2008
VL 26
IS 4
BP 512
EP 524
DI 10.1108/07378830810920860

AU Hine, C
AF Hine, Christine
TI Connective ethnography for the exploration of e-science
SO JOURNAL OF COMPUTER-MEDIATED COMMUNICATION
AB E-science comprises diverse sites, connected in complex and heterogeneous; ways. While ethnography is well established as a way of exploring the detail of the knowledge production process, Some strategic adaptations are prompted by this spatial complexity of e-science, This article describes a study that focused on the biological discipline of systematics, exploring the ways in which use of a variety of information and communication technologies has become a routine part of disciplinary practice. The ethnography combined observation and interviews within systematics institutions with mailing list participation, exploration of web landscapes, and analysis of expectations around information and communications technologies as portrayed in poky documents, Exploring connections among these different activities offers a means to of understanding multiple dimensions of e-science. as a focus of practice and policy. It is important when studying e-science to engage critically with claims about the transformative capacity of new technologies and to adopt methodologies that. remain agnostic in the face of Such claims: A connective approach to ethnography offers considerable promise in this regard.
SN 1083-6101
PD JAN
PY 2007
VL 12
IS 2
AR 14

Appendix B: Original Informed Consent Form for Physicists (Later: Scientists), 2013

Information Provision within the Global Networked Academic Information Environment: Computer Aided Qualitative Data Analysis and Webethnography as Models for Evaluating Information Provision and the Information-related Behaviors of Physicists

I would like to request your cooperation in conducting an ethnographic study for my doctoral dissertation investigating information and collaboration patterns of physicists [NOTE: later expanded to scientists], including interaction with software, open access resources, other colleagues around the world, and library resources.

This study aims to provide a more nuanced understanding of how physicists currently work within the global networked information environment, and findings will contribute to research in library and information science and information systems design, and may be beneficial in proposing a methodological approach for analyzing how libraries might better serve the needs of physicists in various branches of the field.

You have been selected to participate in this study after consultation with other physicists with whom you have collaborated.

If you decide to participate in this study—*your decision is voluntary and you may withdraw at any time during the interview process* and any data the researcher has gathered during the study will be destroyed if you decide to withdraw—you will be as to complete a short pre-interview questionnaire, participate in a virtual (email or Skype) interview and post-interview discussion, and have your computer research/writing/collaboration habits recorded using screen capture software or personally observed. There are no physical, psychological, social, or legal risks involved in this process.

The participant data will be analyzed in conjunction with theoretical analysis and the results will be reported in my doctoral dissertation. I also plan to publish several articles in library and information science and information systems journals. In reporting on the research, the confidentiality of the subjects will be assured. Any information obtained in connection with this study that can be identified with you will remain confidential and will be disclosed only with your permission.

If you have any questions about your rights as a research subject, you may contact my thesis advisor at the Humboldt University of Berlin’s School of Library and Information Science:

Prof. Michael Seadle, PhD
Humboldt-Universität zu Berlin
Institut für Bibliotheks- und Informationswissenschaft
Unter den Linden 6
10099 Berlin
Germany
Tel: +49 (030) 2093-4248
Email: seadle@ibi.hu-berlin.de

This project has been approved by Humboldt University of Berlin's Institute for European Cultural Anthropology Examination and Ethics Committee (decision dated 7.7.2013):

Prof. Dr. Stefan Beck
Philisophische Fakultät I
Prüfungs- und Ethiksausschuss
Zimmer 202
Mohrenstr. 41
10117 Berlin
Germany
+49 30 2093 70847

If you have any questions related to the research study, please contact the investigator:

Stephanie Krueger, Ph.D. Candidate
Berlin School of Library and Information Science
The Humboldt University of Berlin /
Assistant to the Director
Czech National Library of Technology
Technická 6 160 80 Praha 6
Czech Republic
Tel. +420 232 002 545
stephanie.krueger@student.hu-berlin.de

My signature indicates that I have read the information above and decided to participate. I realize that I may withdraw without prejudice at any time after signing this form should I decide to do so. If you require a copy of this consent form, one will be provided to you.

Thank you.

Participant's signature _______________________________ Date:__________________
Investigator's signature _______________________________ Date: __________________

Appendix C: Request for Participation (RFP) for Physicists (Later: Scientists), 2013

Text of RFP which will be sent directly to identified potential participants

Subject: Request to participate in a doctoral study that investigates information and collaboration patterns of physicists (including interaction with software, open access resources, other colleagues around the world, and library resources)

Dear <Name and Academic Title>:

I would like to request your cooperation in conducting a doctoral study investigating information and collaboration patterns of physicists, including interaction with software, open access resources, other colleagues around the world, and library resources.

Objectives of the Study:
This study aims to provide a more nuanced understanding of how physicists currently work within the global networked information environment, and findings will contribute to research in library and information science and information systems design, and may be beneficial in proposing a methodological approach for analyzing how libraries might better serve the needs of physicists in various branches of the field.

Participant Characteristics:
The participants in this study have been selected as leaders in their field, representing different branches of study as well as being located in different geographic regions, and were identified in conjunction with consultation of other physicists with whom you have collaborated.

The Procedure:
As a participant in the study, you will be as to:

- Complete a short pre-interview questionnaire
- Participate in a virtual (email or Skype) interview. Interview questions will be provided to you in advance, and the interview is intended to take approximately 30 minutes to complete. Access to the interview transcript will be available only to the investigator.
- Post-interview discussion, if the investigator needs clarification regarding the interview data.
- Recording your computer research/writing/collaboration habits using screen capture software for one hour over a period of five days and uploading the video capture to an online collaboration space. The investigator will provide instructions for using the software and uploading to the online collaboration space prior to your recording sessions. Access to the videos will be available only to the investigator. Alternatively, the investigator will observe your research habits in person, if I am located near you (i.e., in Prague, Czech Republic).
- Possible other follow-up discussions or as part of this ethnographic study.

There are no physical, psychological, social, or legal risks involved in this process.

When?
Participants will be asked to be interviewed and to record their research/writing/collaboration habits between August 1, 2013 and March 30, 2014.

Outlets for distribution of research results:
The participant data will be analyzed in conjunction with theoretical analysis and the results will be reported in my doctoral dissertation. I also plan to publish several articles in library and information science and information systems journals. In reporting on the research, the confidentiality of the subjects will be assured. Any information obtained in connection with this study that can be identified with you will remain confidential and will be disclosed only with your permissions.

If you would like to participate:
Please contact Stephanie Krueger at the address provided below. Once the time is set for the interview and video recording/in-person observations, you will receive the pre-interview questionnaire, the interview questions, and the consent form that includes information as mandated by my university. I will also talk to you about the process and answer any questions you may have about the study.

Kind regards,

Stephanie Krueger, Ph.D. Candidate
Berlin School of Library and Information Science
The Humboldt University of Berlin /
Assistant to the Director
Czech National Library of Technology
Technická
160 80 Praha 6
Czech Republic
Tel. +420 232 002 545
stephanie.krueger@student.hu-berlin.de

Appendix D: Pre-Interview Questionnaire, 2013

Pre-Interview Questionnaire

Please describe your general area of research.

Do you ever use information resources in your research or to keep aware of publication trends? (library article/journal databases, pre-print websites such as arXiv.org, Wikipedia, repositories, blogs, etc.)

If yes, what are the top five information resources of most importance to you, and why?

What software tools do you commonly use to collaborate with other researchers (e.g., Skype, Dropbox, GoogleDrive, other)?

What software tools do you commonly use in your research (e.g., Surface Evolver, http://www.susqu.edu/brakke/evolver/evolver.html, etc.)?

For writing publications, do you use citation management tools (e.g., EndNote, ProCite, RefWorks, Zotero)?

Have you ever experienced library services that have helped you in your research or education? If so, can you describe the services?

Do you think the library as an institution has any place for scholars in your field?

Appendix E: Sample Joint Request for Participation with Research Participant, 2013

Text distributed by Participant 1 via email (n= 5) to community of selected theoretical physicists.

Dear X,

I am working with PhD candidate Stephanie Krueger, who is studying the collaboration methods of physicists to better understand how ideas are ultimately transformed into results over global communication networks. For instance, how does a physicist's use of the Internet to do research and communicate affect the "paper-building process." To conduct the study, Stephanie needs subjects, i.e. physicists who are open about their research habits. Participation in the study would not involve much more than filling out a questionnaire, an interview or two (with you and potentially members of your group), and some recording of your research habits via screen captures.

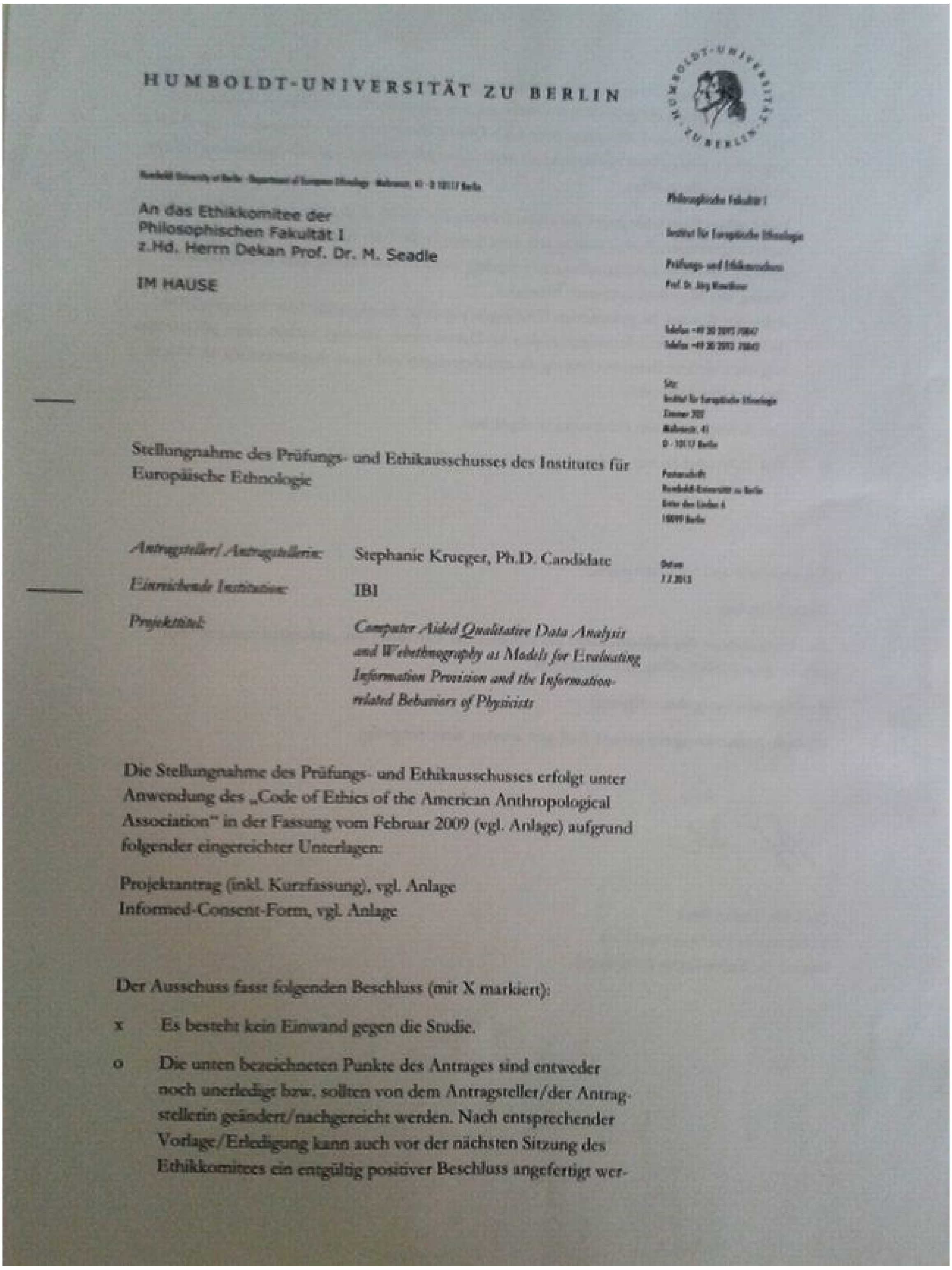
All of this info would remain confidential and may lead to some understanding of how the Internet can streamline the paper-building process. On the other hand, open participation could presumably be used as part of your outreach/education platform.

Should you be interested in participating in this study, a letter from Stephanie is attached with a few more details. Her university email is above in the cc field if you want to contact her directly. Also, I would be happy to talk with you about this study via phone/e-mail/Skype. My phone number is 607-342-0876, and Stephanie's university email is above in the cc field.

Finally, please forward this e-mail to physicists you know who might be interested in participating in this kind of study.

Cheers, X
(sent with Informed Consent Form above)

Appendix F: Humboldt Project Approval, 2013



HUMBOLDT-UNIVERSITÄT ZU BERLIN

Humboldt University of Berlin · Department of European Ethnology · Mohrenstr. 41 · D 10117 Berlin

An das Ethikkomitee der
Philosophischen Fakultät I
z.Hd. Herrn Dekan Prof. Dr. M. Seadle

IM HAUSE

Philosophische Fakultät I
Institut für Europäische Ethnologie
Prüfungs- und Ethikausschuss
Prof. Dr. Jörg Niewöhner

Telefon +49 30 2093 70847
Telefax +49 30 2093 70843

Sitz:
Institut für Europäische Ethnologie
Zimmer 207
Mohrenstr. 41
D - 10117 Berlin

Postanschrift:
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

Datum
7.7.2013

Stellungnahme des Prüfungs- und Ethikausschusses des Institutes für Europäische Ethnologie

Antragsteller/Antragstellerin:	Stephanie Krueger, Ph.D. Candidate
Einreichende Institution:	IBI
Projekttitel:	*Computer Aided Qualitative Data Analysis and Webethnography as Models for Evaluating Information Provision and the Information-related Behaviors of Physicists*

Die Stellungnahme des Prüfungs- und Ethikausschusses erfolgt unter Anwendung des „Code of Ethics of the American Anthropological Association" in der Fassung vom Februar 2009 (vgl. Anlage) aufgrund folgender eingereichter Unterlagen:

Projektantrag (inkl. Kurzfassung), vgl. Anlage
Informed-Consent-Form, vgl. Anlage

Der Ausschuss fasst folgenden Beschluss (mit X markiert):

x Es besteht kein Einwand gegen die Studie.

o Die unten bezeichneten Punkte des Antrages sind entweder noch unerledigt bzw. sollten von dem Antragsteller/der Antragstellerin geändert/nachgereicht werden. Nach entsprechender Vorlage/Erledigung kann auch vor der nächsten Sitzung des Ethikkomitees ein entgültig positiver Beschluss angefertigt werden. Der Antrag wird in der nächsten Sitzung der Kommission nicht mehr behandelt. Achtung: Werden die geforderten Unterlagen von dem Antragsteller/der Antragstellerin nicht innerhalb von 3 Sitzungsperioden (ab Datum dieser Sitzung) nachgereicht, gilt der Antrag ohne weitere Benachrichtigung als zurückgezogen und muss gegebenenfalls als Neuantrag eingereicht werden.

- o Es bestehen Einwände gegen die Durchführung der Studie in der eingereichten Form. Die unten angeführten Punkte sollten von dem Antragsteller/der Antragstellerin entsprechend geändert und dem Ethikkomitee neu vorgelegt werden. Der Antrag wird in der nächsten Sitzung des Komitees nochmals behandelt.
  Achtung: Werden die geforderten Unterlagen von dem Antragsteller/der Antragstellerin nicht innerhalb von 3 Sitzungsperioden (ab Datum dieser Sitzung) nachgereicht, gilt der Antrag ohne weitere Benachrichtigung als zurückgezogen und muss gegebenenfalls als Neuantrag eingereicht werden.
- o Der Antrag wird vom Ethikkomitee abgelehnt.
- o Der TOP wird bis zur nächsten Sitzung vertagt. (Begründung siehe unten)

Kommentare und Empfehlungen:

Zum Prüfplan: –

Zur Information der Informanten: das vorgesehene Verfahren des „informed consent" entspricht allen zu stellenden Anforderungen.

Zur Versicherungsbestätigung: –

Weitere Anmerkungen: weitere Auflagen werden nicht festgelegt.

[redacted]

Prof. Dr. Stefan Beck
Vorsitzender Prüfungsausschuss
Institut für Europäische Ethnologie

Appendix G: Visualization Background and Descriptions

G.1 Graphing and Visualization Tool: Wolfram Mathematica 10.2

I created all graphs and visualizations in this thesis using a trial version of Wolfram Mathematica 10.2, having heard about the tool from Participant 1 and also having discussed it with a professor in Mathematics at Czech Technical University in Prague
(S. Krueger 2015, pers. comm. with mathematics professor, 20 August):

> *Dear Stephanie,*
>
> *personally, I would like very much the 3D graph in Mathematica on slide 2, use 2 colors for the 2 types of connections (edges), and highlight the stronger connections by thicker lines; a very basic example is attached. Unfortunately, I do not know any software that is better than Mathematica for this purpose, but I will try to find out more.*

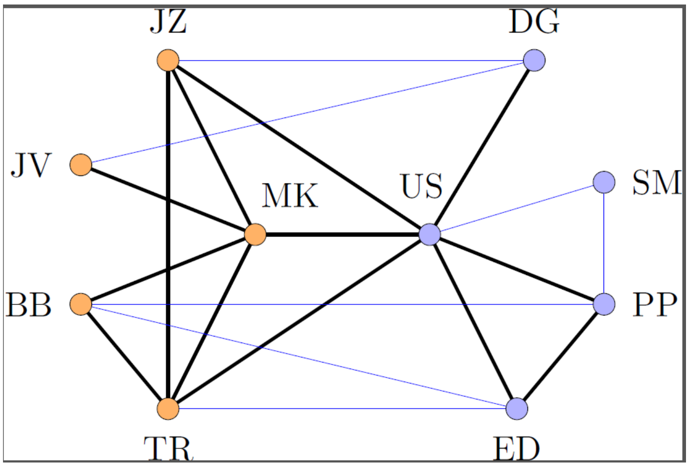


Figure 38: Network Graph Illustration from CTU Prague Mathematics Professor (attachment to email above)

I considered using open source KNIME for creating graphs and visualizations, but found the graphics generated by Mathematica to be more visually appealing.

G.2 Visualization Type One: BubbleChart3D

I utilized Mathematica's BubbleChart3D to create the visualizations of individual resource use. In these charts, each resource is represented by a colored "bubble" (one bubble=one resource) and placed in a matrix according to coordinates in the vicinity of other similar resources (e.g., non-profit resources that do not require library intermediation for access to them). The following figure illustrates how one of these charts was created, for readers interested in this topic.

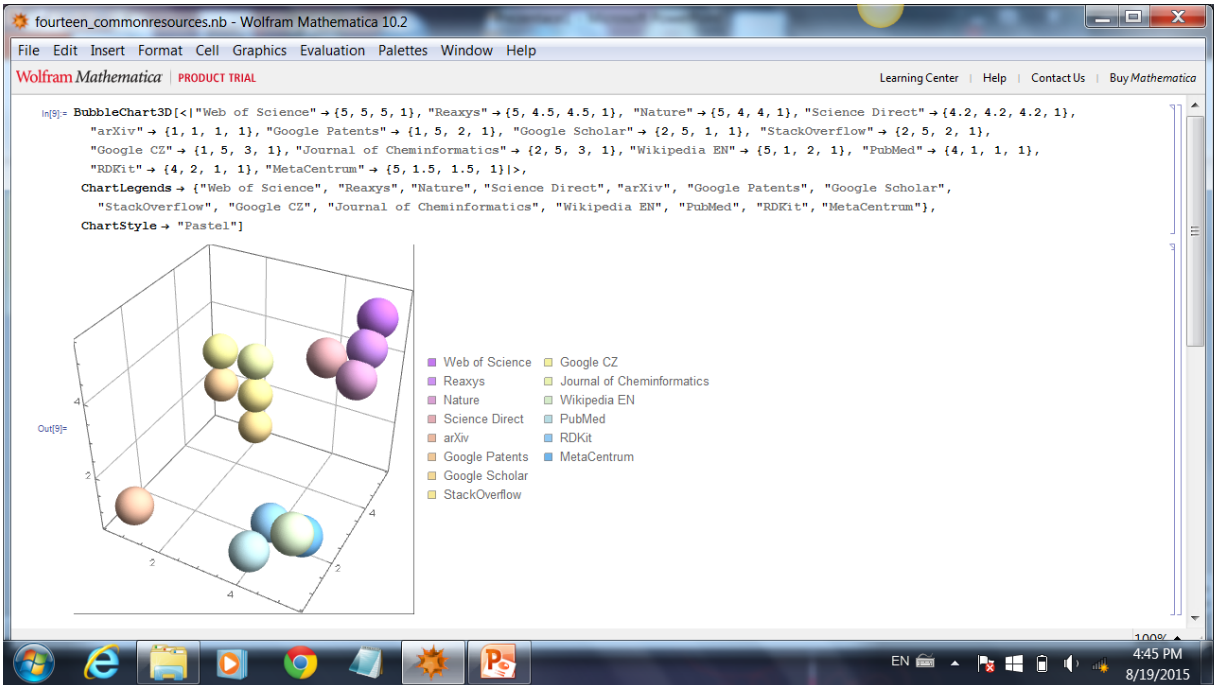


Figure 39: Mathematica Interface for BubbleChart3D.

G.3 Visualization Type Two: Network Graph

I used network graphs, as in the example from the mathematics professor above, to illustrate interactions of participants with individual information resources or groups thereof. To do this, I worked with my raw fieldwork and screenshot resource data and then defined the relationships between entities—I assigned both participants and resources in this case alphabetic identifiers, which are visible in the following figures.

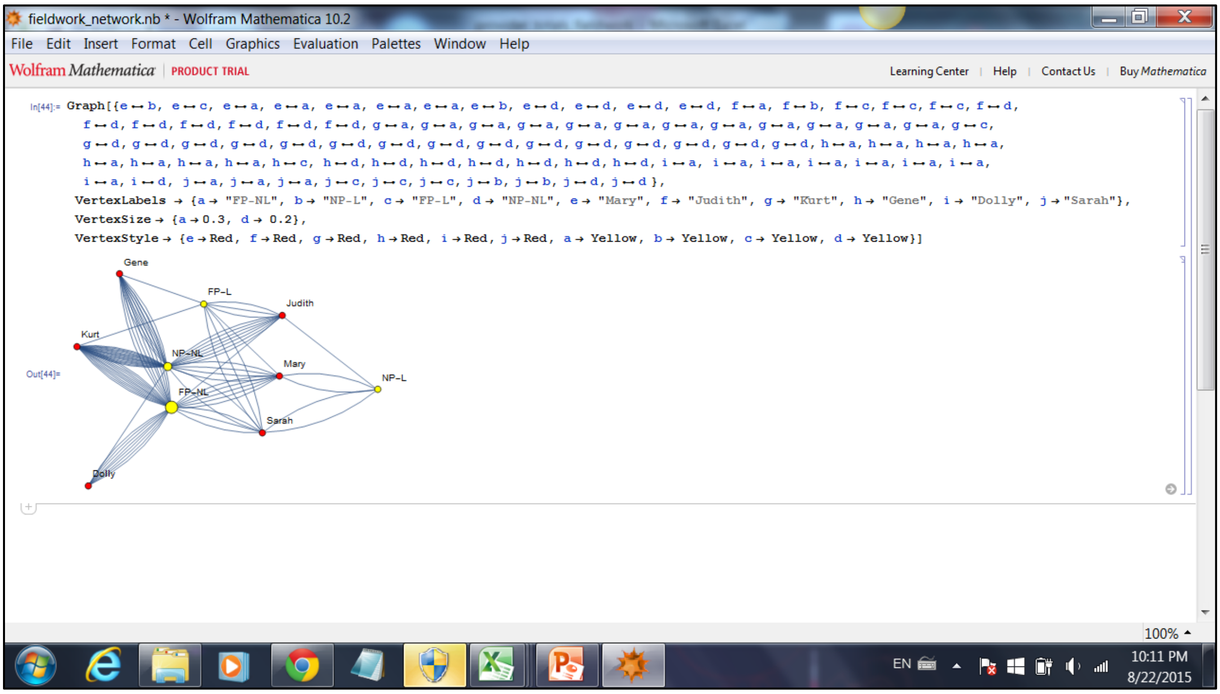


Figure 40: Mathematica Interface for Network Graph of Fieldwork Data, Information Resources Grouped.

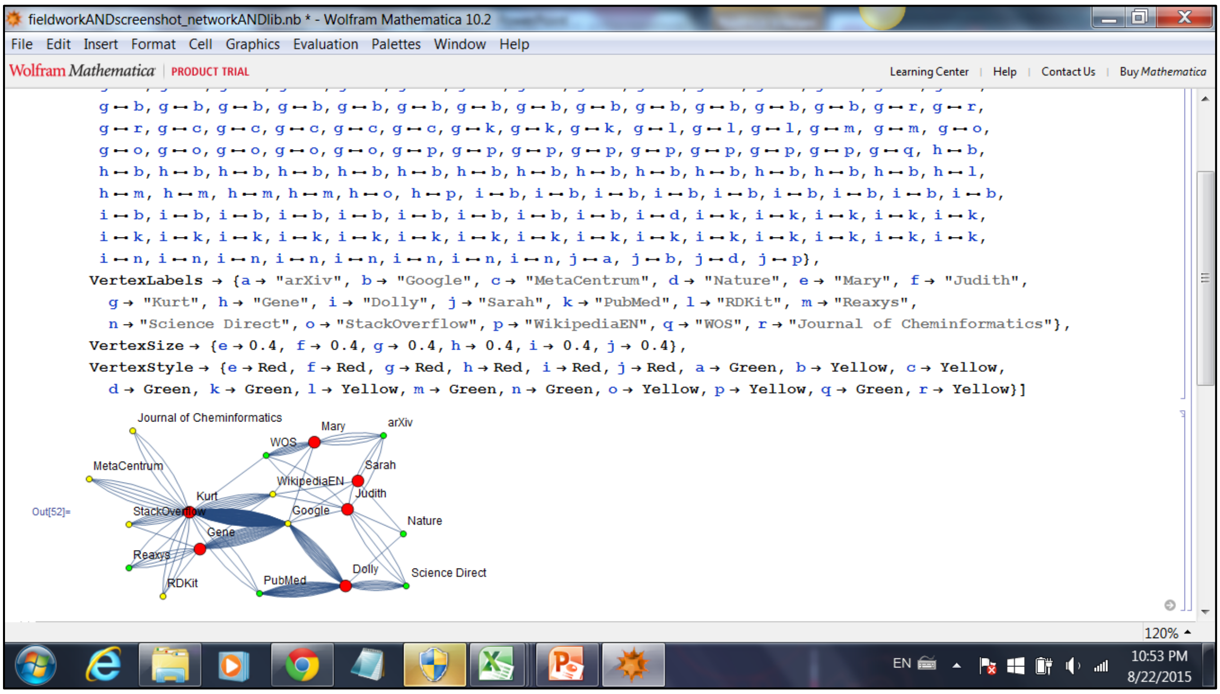


Figure 41: Mathematica Network Graph, Individual Resources and Participants.

Full input sequence for the graph above:

```
Graph[{e<->a,e<->a,e<->a,e<->a,e<->b,e<->p, e<->q,e<->q,e<->q,e<->q,e<->q,e<->q,e<-
>q,f <->a, f<->a, f<->b, f<->d, f<->n, f<->n, f<->n, f<->p, f<->q, f<->q, g<->b,g<->b,g<-
>b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<-
>b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<-
>b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<-
>b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<-
>b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<-
>b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->b,g<->r,g<->r,g<->r,g<->c,g<->c,g<->c,g<->c,g<-
>c,g<->k,g<->k,g<->k, g<->l,g<->l,g<->l,g<->m, g<->m, g<->o,g<->o,g<->o,g<->o,g<->o,g<-
>o,g<->p,g<->p,g<->p,g<->p,g<->p,g<->p,g<->p,g<->p,g<->q, h<->b,h<->b,h<->b,h<->b,h<-
>b,h<->b,h<->b,h<->b,h<->b,h<->b,h<->b,h<->b,h<->b,h<->b,h<->b,h<->l,h<->m, h<->m,
h<->m,h<->m,h<->o, h<->p, i<->b,i<->b,i<->b,i<->b,i<->b,i<->b,i<->b,i<->b,i<->b,i<->b,i<-
>b,i<->b,i<->b,i<->b,i<->b,i<->b,i<->b,i<->b,i<->d,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<-
>k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->k,i<->n,i<->n,i<-
>n,i<->n,i<->n,i<->n,i<->n,i<->n,j<->a, j<->b, j<->d, j<->p},VertexLabels -> {a-> "arXiv", b->
"Google", c-> "MetaCentrum", d-> "Nature", e-> "Mary", f-> "Judith", g-> "Kurt", h->
"Gene", i-> "Dolly", j->"Sarah", k->"PubMed",l->"RDKit", m->"Reaxys",n->"Science
Direct",o->"StackOverflow",p->"WikipediaEN",q->"WOS",r->"Journal of
Cheminformatics"}, VertexSize-> {e->0.4, f-> 0.4,g-> 0.4,h-> 0.4,i-> 0.4,j-> 0.4},
VertexStyle-> {e->Red, f->Red, g->Red, h->Red, i->Red,j->Red, a-> Green, b-> Yellow, c->
Yellow, d-> Green, k-> Green,l-> Yellow,m-> Green,n-> Green,o-> Yellow,p-> Yellow,q->
Green,r-> Yellow}]
```

Appendix H: Kurt DeSilva Screenshot Commentaries, 2015

21. Format: In-person with p5 plus email follow-up
Participant: p4
Date: 29 July 2015
Storage Location: mail.techlib.cz
Theme: update their work
Excerpt (text or image):
here is the clarification of the screenshots:
120455: StackOverflow search for how to force figure placement in text in LaTeX (was writing an internal grant application at the time, I think)
170750: linux documentation manpage - resolving software dependencies
170803: same as 170750, sorry for the duplicity
163005: requested screenshot - passing rates for math I (mandatory freshman course)
163419: the same, but for chem I
160338: a quick fulltext search for '::mol' in RDKit database cartridge documentation - was just looking up the method for loading a SMILES string as a RDKit structure
162447: the same as 160338, fulltext search for 'morgan', as just quickly wanted to get a method for morgan fingerprinting of RDKit structures
Analytic Commentary/CODES:
RES TOPIC

22. Format: email
Participant: p4
Date: 8 August 2015
Storage location: mail.techlib.cz
Theme: More context for screenshots
Excerpt:
174932: I think that's the very first screenshot after I've installed the 'scrot' screenshot program (output of the install process is the wall of text in the console on the right). After installation, I tried to run the command (the last row of the console) and it snapped this screenshot of the whole desktop.
175332: I had a pharma assignment to find an industrially viable way of producing the 'rolapitant' drug from available info sources. A part of rolapitant making process is olefin metathesis, which uses ruthenium-based catalysts. Ruthenium is quite toxic, so I was searching if I can get it out of the reaction mixture after it has done its job.
175407: The same project, I was just reading up on the whole metathesis thing.
175426: Answer to the Ruthenium removal thing was found on the 'All things Metathesis' blog. Informal, but very informative.
175453: A search for the catalyst in question. Is it available?
175527: Turns out it is available, right at Sigma-Aldritch. Over 10.000 CZK per gram (that might be even more expensive than coke), but available, and that's all that matters for the scope of my work.
175707: The original patent describing, among many other things, a working (but just lab-scale) rolapitant-making process. This was the most informative source by far, but also a bit annoying to read due to referencing many other papers, convoluted language and back-and-forth document self-references.

180746: I wasn't sure what the 'celite pad' mentioned in the patent was, so I googled it and facepalmed the moment I saw the pictures. Yes, I am theorist and it really shows :)
183623: The same as above, just search for 'brine'. Turns out it's just a fancy word for salty water.
184322: Another step in rolapitant prep is hydrogenation catalyzed by Pd on C. I was already somewhat familiar with this, but googled just to make sure.
131959: Was searching for this compound mentioned in the patent
120431: I needed to place an unruly figure in my rolapitant report that I was typesetting with LaTeX.
120455: Answer to the above - StackOverflow to the rescue
121255: Seach for czech sources on correct manufacturing practice. It was again rolapitant-related, but I can't remember how.
132007: LaTeX kept breaking my lines where I didn't want them to. I needed to break manually due to the ofter line-or-longer chemical names.
133117: Direct search for a patent by patent number, again rolapitant-related. Google is just fine for that.
133238: Generic Python programming query, seems like a stupid one, too. Can't remember details, sorry.
133840: Another search for a compound mentioned in the rolapitant-related patent.
133858: Found it at Sigma-Aldritch, easy to get and affordable at 10 CZK a gram.
134048: Didn't know what 'chiralcel separation' is. Seems to be a proprietary lab method of separating isomers.
103237: Can't remember why I searched for that
142711: Grabbing XML publication data from Reaxys for a czech chemist for our little CzeChem project.
144924: Same CzeChem project as above, searching for the Vladimir Sindelar chemist so I can export all compounds published in his works.
154526: Generic linux stuff, about generating checksums to validate my program exports, I think.
174740: PostgreSQL (~ postgres) generic query to make sure I understand the implementation of certain data types properly.
174821: The info that I got from a hit from the previous query. Just postgres documentation. I learned that for me it is much quicker to just google specific things than going through the content tree of documentation.
181136: The same postgres stuff, this time about granting permissions so that my scripts can access the database, but not overwrite primary data. Its better to make sure at the lowest possible level.
170654: This was a bit dumb query, I was hoping for some miracle way of porting my heavy RDKit toolkit. Nope.
170735: Looking for how to use the 'ldd' linux program that tells you programming library dependencies of a given piece of software.
170750: Just landed on the standard manual page for the 'ldd'. Could have just typed 'man ldd' instead of googling.
170803: Snapped the same thing twice. Sorry for that.
130923: Django is a Python framework that I use to provide some sort of user interaction and handling more complex processing over my chemical databases. Now I was search for a way to reset the database (to clear it), but not to drop/delete the tables themselves.

132240: git is a versioning system for (mainly) software project that keeps track who changed what, when and where in the code. I wanted to get the view side-by-side of two versions with highlighted inserts and deletes. I know it's there, but I couldn't remember the command, probably.
171131: Searched whether MetaCentrum where I compute most of my stuff supports Numpy (a python library for advanced computational/statistic things)
110018: More Django stuff, wanted to insert some data into a HTML request. Can't remember exactly why.
125336: The time-honored tradition of copying an encountered programming error into google, verbatim :) It usually works.
115021: I was searching for alternative libraries for python<-->postgres communication. I already know and use the standard psycopg2, but if I remember correctly, I wanted something more portable.
145253: Generic linux stuff, was searching for a way to easily blacklist some advertising IPs. Not really work-related.
163005: Requested screenshot of the abysmal pass rates for our freshman Chem101.
163419: The same as above, but for Math101.
171812: Probably some Postgres configuration trouble related query, can't remember this one
175900: Same thing as when I searched in 115021, I wanted something portable yet psycopg-like.
190529: A search for a way to compress a Python project into one easy-to-move executable via cx_freeze. Later decided that it's too much effort for too little/inconsistent reward.
161023: Don't remember this particular reference.
172428: Looking up a drug group, probably gathering info for my pharma exam.
140829: Studying for pharma exam, I think.
131116: I don't know why I snapped the entire desktop... but the query on the left monitor is about python API (application programming interface) for MySQL.
131142: The same as above, this time I managed to hit :)
151002: Searching about search implementation in the Django framework to better access my chemical DBs.
154335: The same as above, this time about searching over specific data fields (boolean). I probably wanted to do some cheap screening (throwing away compounds that can't possibly match the query, like for example if somebody searches for a ringed substructure, I can go ahead and throw away anything that doesn't have a ring(easy check) before trying to fit the substructure in (computationally much more expensive check))
110201: 'Promiscuous receptor' sounded a bit funny in my mind, so I googled it to make sure that it's the correct technical term.
165622: A query about adding new identifier columns to an existing postgres database table. When I alter already filled tables, I usually check and double-check before putting in any commands. A fatal crash in any useable system is always exactly one command away :)
165720: Dumping a table is a way of data backup into (usually) text formats. Related to previous query. Doing backups before running the command.
131214: Colleagues wanted some data from me, so I was searching for a way to export a csv file directly from the dabase. Luckily it's quite easy...
131231: ... as seen on this StackOverflow screenshot.

135901: I wanted to list all postgres functions available from a given database, probably to find out whether my RDKit database cartridge functionality was detected in full.
135925: ... answer again found on StackOverflow
141251: I decided to write some postgres database functions of my own, to move as much functionality as possible into the database itself (if data are smart, the program using them can be much dumber)
141902: Related to previous query, googling for the very basics of postgres functions. I know what I want to do, but don't know the postgres programming syntax.
143544: Again, postgres syntax stuff
151645: Moving up a bit from the very basics, learning how to program a transaction (either all requested things get done or none do, nothing in between. Useful for data integrity).
160241: More postgres programming basics
160315: Adding the RDKit cartridge functionality into previously purely postgres functionality
160338: Consulting the RDKit cartridge manual for that function, just punched '::mol' into fulltext search
162247: After succesful structure import, I search for in-database fingerprinting methods
162447: ... and find them again the RDKit cartridge manual via fulltext search
163313: Postgres didn't like my new code too much, so again, googling the error verbatim
111408: Refreshing my memory on the database cursor for the psycopg2 python library
162043: Resolving the script access problems - postgres needed to be explicitly told to allow the script to search over certain tables
164842: Looking for a way to make directories using a core os Python library. Elementary stuff, but it's the first time that I actually had to use it.
170234: I was wondering why I couldn't make any permanent changes in the database from the script. Turns out, I wasn't commiting the changes manually and the autocommit was off.
110600: Posgres also needs to manage permissions for sequences, not just the table. If a script wants to insert a new row into a table that he has the relevent permissions on, it still doesn't imply that in can access the sequence that generates unique ID's for the rows. StackOverflow saved me a lot of grief once again.
112825: Another verbatim googling of postgres error.
113316: Learning how to gracefully handle empty array in postgres.
150958: I've apparently wanted to download some REACH chemical substances data sheets, but I can't remember why. Probably for my pharma project.
151117: I don't remember why I wanted general tox data at the time, sorry.
162251: 'physical-chemical methods' just sounded silly in my mind, so I googled it to make sure it's used and established expression before I put it into my writing :)
162254: I wanted to do an aggregate median in postgres, and it turned out that it isn't directly implemented. The workarounds I found in google hits all seemed kind of messy, so I just implemented it at postprocessing in the python scripts.
135549: Looking up a way to generat MACCS fingerprints by RDKit while being too lazy to dig through the whole RDKit documentation. The third hit is what I've looked for.
135608: Looking up viable ways of randomly sampling database of millions of compounds.
160656: I wanted to know more about the 'qsub' command used by MetaCentrum to queue up computation requests, as I wanted to specify some environment variables for my scripts to lean on.
161434: Related to the previous query, I wanted a way for the scripts to test whether all required variables are set, and perhaps have some sane defaults.

102353: This is my most read cheminformatic source -- the journal of cheminformatics. It's open-access, so I simply have my RSS reader set to have it as a source among other news. I usually read all the articles there the day they come up. Just click and read, no hassle with logging in and verification.
154315: A good portable python<-->postgres library has finally been found (by Milan), but it doesn't support user-defined types (like fingerprints). Meh.
141537: I was working with some postgres enum types (enum=pick one and only one value from a list of values. An example would be a day in a week: one of [Mo, Tu, Wed...]). I was searching for a way to list all available values for a given enum in postgres.
164048: More MetaCentrum stuff, I was researching a specific computation request parameter, but I can't remember why.
164214: No clue what that is, probably a MetaCentrum layer thing.
093005: Reading up on GiST indexes in postgres. Searching for a way to speed up fingerprint searches.
173703: An overall screenshot of my MetaCentrum computation autoqueueing script (running on the left console) and MetaCentrum account run data.
173727: A cutoff from the previous view. Please note the amount of user processorhours :)
122108: Just a screenshot of 244 computer cores running at the same time. No science, just a silly nerd cred thing...
141232: Same as above, just even more cores.
100740: I had multiple databases side-by-side, and was wondering whether it's possible to make a query spanning more databases. From StackOverflow - it is possible, but with many issues I can't be bothered to address.
102544: Here I am not sure, but I think I was wondering whether I can shorten in postgres the notation of 'if thing_to_look_for=True; then do_stuff' into a more elegant 'if thing_to_look_for; then do_stuff'
133231: A new article in Journal of Cheminformatics just popped up. Same context as 102353.
103115: Same as 141232, with many, many cores.
172055: Search related to our current project, running Molpher on Glucocorticoid receptor agonists. I was searching to see if there are any already established methods to screen for GR ligands.
172108: Related to the previous query. Found & read this article.
172122: Another relevant hit to the previous query.
103201: Same as 103115, even more cores.
152817: I wasn't sure what 'cocrystal' is, so I spent some time on wikipedia repairing this hole in my education. It probably came up in the context of protein--ligand interaction modelling.
152827: Zoom-in from the previous screenshot.
154726: I am not sure why I searched for that.
134821: StackOverflow, a simple, to-the-point answer to a simple, specific question on how to select table rows in postgres that are not in the other table. (usually it's the other way around, so this is a bit of a special case).
102513: I think Dan mentioned Tropha's choices of descriptors for Quantitative Structure-Activity Relationship models, so I had a look myself.
134454: An embarassingly simple query on postgres features. I remembered natural join from before, but wasn't sure about the specifics in naming of the columns to join over.

160126: Query related to improving postgres performance by partitioning the data. Most of it went over my head.
142108: Reading up on transcription factors in general. Related to our favorite glucocorticoid receptor, which is a significant transcription factor.
142117: Reading up on the glucocorticoid receptor itself.
142130: The same as above, but on adrenal cortex.
142211: Following wiki link into some actual scientific literature.
142503: Reviewing some of the known and used glucocorticoid receptor agonists: Dexamethazone. Note the steroid skelet of the compound... we want new and non-steroid stuff.
150157: Searching for known glucocorticoid receptor ligands in the ChEMBL database.
163130: Looking for what happends to mammals without working glucocorticoid receptor.
173015: Finding more ammo for my case why new glucocorticoid receptor agonists are needed.
173027: The same as above without the sloppy borders.
130857: More research into glucocorticoid receptor - related aiments.

Appendix I: Pilot Instruction for Czech Technical University in Prague (CTU Prague) Based on Research Findings, 2014

(S. Krueger 2014, pers. comm., 11 Sept)

Dear all,

Brief notes from today - please add comments if I missed something or misunderstood something we discussed.

All my best - looking forward to our pilot!

Stephanie

***
Classes - Fall/Winter 2014/15 Pilot with NTK

Primary Goals of the Pilot:
1. **Reduce** basic library/information resource teaching burden for you and your colleagues
2. Provide "**QuickStart**" to effectively using eResources, eBooks, and other materials provided by NTK and CVUT and openly online - advanced skills for serious young scholars
3. Give an **introduction to the scholarly publishing/research** process/lifecycle
4. Determine what **additional advanced instruction or services NTK** might provide which are not available from the CVUT library presently
5. Give students an opportunity to practice their **research English** - oral and written skills

Instructor: Mgr. a Mgr. Stephanie Krueger (you have my LinkedIn bio)
Language of class and assignments: English

Format:
- Four **30 minute classes at NTK and the department** once a month in Oct, Nov, Dec 2014 and Jan 2015
- Four **optional** brief written homeworks will be assigned and corrected by the instructor (me, native English speaker); instructor will give the assignment back to students for their review at each class and provide time beyond class to discuss questions, comments in more detail upon demand
- No grades this time

Oct class (at NTK)
Title: <u>Beyond Google: QuickStart to effectively using NTK and CVUT resources</u>

Includes remote access, special services for PhD students, building a good query in English
Homework: Write a one-page description of your research interests and prior experience with library/information resources

Nov class (at Faculty of Civil Engineering)
Title: The scholarly publishing/research universe: what you might not know yet & tools to help you

Includes overview of citation management options (open source and commercial - review your existing guide briefly at: http://en.wikipedia.org/wiki/Comparison_of_reference_management_software + AMS TeX resources, http://www.ams.org/publications/authors/tex/tex), new social tools like ResearchGate, Mendeley - advantages and disadvantages, effectively using Scopus & Web of Science, critical thinking in relation to commercially-provided tools (pros & cons). Also, brief discussion of tracking **current contents/effective use of RSS** (what is it, etc.)

Homework: Write a brief (one paragraph) description of articles you have already published or describing proposals for conferences to date, then write two one paragraph reviews discussing the advantages and disadvantages of ResearchGate versus Mendeley and Scopus versus Web of Science, respectively

Dec class (at Faculty of Civil Engineering)
Title: Professional scientific writing for civil engineers: useful resources and tips

Overview of useful specialized reference materials, the style of scientific writing, where to look for help, submission guidelines - what are they, what to consider

Homework: Write a one-page review of your favorite print or electronic resource(s) for civil engineering - choose one or discuss several - not more than three favorites. What do you like and *what drives you crazy*?

Jan class (at Faculty of Civil Engineering)
Title: Evaluating citation metrics, journal impact factor, and the world of the scientific author (your rights as an author, what you need to know about open access, copyright generally)

The world of citation metrics, impact factor, and scientific author rights is increasingly complex. We will discuss effectively navigating this complex jungle as well as critical evaluation of various rating schemes and mechanisms.

Homework: Find a **sample author contract** for a journal you publish in OR would like to publish in someday. Write a one-paragraph summary of your rights as an author, and a one-paragraph statement about whether or not you like these author's rights.

***
Instructor will send homework comments back from this last class back via email, and will conduct a **brief course evaluation** so that all pilot participants can refine future course ideas.

Appendix J: Proposal, Introducing Scientific Writing into the Engineering Curriculum, 2015

Introducing Scientific Writing into the Engineering Curriculum

Organized by the National Library of Technology (NTK) for PhD students and researchers at CTU in Prague

Executive Summary

This proposal supports the introduction of university-funded scientific writing in English courses to all interested doctoral students as well as junior and senior researchers across all CTU faculties/departments. These courses, one semester in length and offered at the centrally-located National Library of Technology, would provide a critical foundation for fostering the global competitiveness of CTU's most promising young scholars, giving them the confidence and skills to publish in the highest-quality international journals and to masterfully present at international conferences—activities which, over the long-term, will contribute significantly to CTU's domestic and global rankings. University investment in these courses would be primarily for creating half-time instructor positions—during stage one (the project pilot), one half-time instructor would be required; in stage two (ongoing operations), an additional half-time position would be needed in order to enable additional courses as well an English editorial service. Participation in the courses would be voluntary. The selection process would be managed centrally by the instructor(s) plus NTK staff (as needed); each CTU faculty, through its Vice-Dean of Research, would receive a certain number of places in the courses for its staff or doctoral students, determined according to scientific output of faculty measured by RIV points. Target start date for stage one: September 2015; stage two: September 2016.

Background

The most direct way to improve the quality of research and to foster the global competitiveness of CTU is to encourage PhD students to **publish** in **high-quality journals**. Our doctoral students and young researchers – in comparison to those at the world's top-rated engineering universities – are hindered in their academic careers by **limited writing skills** and **presentation abilities**, since these are often not part of their undergraduate training and experience. These limitations are sometimes overcome in the course of PhD training, at the expense of an **ad-hoc**, and rather **time-consuming**, involvement of supervisors and/or junior and senior colleagues. The aim of this course is to provide doctoral students (and CTU researchers) with a **systematic** and **professional** education in the area of scientific writing in English. The course would take place and be organized by NTK, due to its central campus location and ability to organize high-quality instruction in the English language. The course would be open to all CTU faculties/departments.

Audience

Stage one (target start date: winter semester 2015/16 with another cycle in spring semester 2016; July-September 2016=planning phase for stage two): classes and services open to **3+ year PhD students, junior and senior researchers**; stage two (2016 and beyond): classes and services open to **first- or second-year PhD students**.

- For all audiences: twelve weeks of instruction (one class per week for one semester, followed by exam in week thirteen), corresponding to the academic calendar.

Proposed Course Content

Stage one course content[1]	Stage two course content
1. Essentials of grammar -- paragraph structure, word order, articles, punctuation, etc. 2. How to organize the writing process and to effectively use library resources 3. Title, abstract, and introduction 4. Materials and methods 5. Writing up research: results and discussion (including data commentary) 6. Conclusions 7. Citations, list of references, and acknowledgements 8. Submission process, reviewing for journals 9. Typesetting of scientific papers (mathematical symbols, units, LaTeX) 10. Presenting conference papers 11. Professional correspondence: emails, CVs, reports 12. Meeting grant and open access requirements: repositories and ranking systems 13. Examination	1. Essentials of grammar -- paragraph structure, word order, articles, punctuation, etc. 2. How to organize the thesis writing process 3. Effectively using library resources 4. Title, abstract, and introduction 5. Materials and methods 6. Writing up research: Results and discussion (including data commentary) and typesetting of scientific papers (mathematical symbols, units, LaTeX) 7. Conclusions 8. Citations, list of references, and acknowledgements 9. Organizing the text of your thesis (proper dissertation format) 10. Ethical issues of dissertation and scientific writing 11. Presenting conference papers 12. Professional correspondence: emails, CVs, reports 13. Examination

Course Capacity and Format

Each course should be limited to 15 students to ensure high interaction between the learners and the instructor.

- Class length: 60 minutes/week with 30 minutes after class for individual consultation and activities.

Ideally in stage one, four sections of the class (60 "seats" for students/researchers per semester, with one 0,5 FTE instructor plus guest speakers as needed for library resource and citation management instruction) would be offered per semester—meaning an instructor would teach four ninety-minute classes (i.e., 60 minutes of lecture followed by 30 minutes for individual consultation and hands-on activities) per week.

- In stage two, additional sections should be added based on demand identified in the first phase; the goal would be to offer at least eight sections (eight classes per week; 120 "seats" per semester with 2x0,5 FTE instructors).

The instructor(s) would be available in the library (office or in a public space whenever classes are not in session) for individual consultations, professional article/book editing in English (a service open and publicized to the entire CTU community, not just those enrolled in classes), class preparation (approximately one hour/week per class in stage two), homework grading (approximately two hours/week per class) and drop-in consultations. The instructor would also be "on-call" for the CTU PhD and researcher community, scheduling consultations in researchers' offices or online so that members of the CTU community would not need to go to NTK in person.

- Note: **capacity for individual editing and consultation would be limited to approximately four hours** for a half-time appointee during teaching semesters in stage two, with more time for editing and consultations becoming available in the examination periods and vacations (NTK staff work twelve months/year). An **extensive editing service is beyond the scope of this proposal**.

Stage two: workload breakdown per 0,5 FTE/week *(assumption: four classes/week. In stage one, class preparation and administrative duties/planning would encompass more time than during later "normal" operations, with class*

[1] Compiled from Effective scientific writing for PhD students taught at VŠCHT Prague; Writing in the Sciences, a Coursera course from Stanford University; and Advanced Workshop on Writing for Science and Engineering and Communicating in Technical Organizations, MIT OpenCourseware seminars.

preparation 6 hours and administration duties 4 hours, leaving little time for editing and consultations during the teaching periods.)

In-class teaching (regularly-scheduled classes; 90 min/class)	6 hours
Class preparation (one hour/class)	4 hours
Grading of homework (one hour/class)	4 hours
Administrative duties (planning for phase two, work with departments for regular classes)	2 hours
Editing and consultations	4 hours
Total:	**20 hours/week**

Selection Process for Class Participants

In order to properly serve the entire university, it is essential that the process for selecting class participants be clearly defined. Each CTU faculty should receive a certain number of places for its staff or doctoral students, determined according to scientific output of faculty measured by RIV points (in stage two, the number of PhD students in a faculty could be additionally taken into account). Each faculty, through its Vice-Dean for Research, would select the participants based on their own criteria and send the names of participants to the course instructor by a particular date. If any seats in the course remain (during stage one, it will still be unclear what the demand might be), the course instructor and NTK instruction staff would hold a public lottery at a specific date/time prior to the course for any remaining seats.

Comparable Campus Offerings

While preparing this proposal, only one comparable campus initiative was identified to date: the Masaryk Institute of Advanced Studies (MIAS) will be providing one-semester English for Specific Purposes and English for Academic Purposes courses. Students in the MIAS courses will be charged 11 800 CZH including taxes for each course.

Our planned course(s) would benefit talented CTU PhD students by providing them highly-specialized (i.e., specialized scientific, engineering-focused publication training), high-caliber instruction for no additional fee (11 800 CZK is a significant financial hurdle for many students). Our proposed course(s) would be held in the center of campus, and because of NTK's opening hours, evening courses could be offered if they make sense for students (specific times/dates of courses would be evaluated during the stage one pilot).

The CTU Central Library additionally provides a course for doctoral students, but this library initiative is focused on proper use of library resources rather than on scientific, engineering-focused writing instruction and practice. Both NTK and CTU Central Library services would be promoted to course participants to supplement their writing coursework.

Estimated costs

If an agreement is met between CTU and NTK for jointly offering these seminars, NTK could dedicate a classroom and office space to the instructor(s), resulting in no rental fee for the combined initiative. NTK staff would be made available for assisting the instructor with the logistics of reserving the classroom. If possible, it would be helpful for CTU to give NTK permission to email Vice-Deans of Research on behalf of the instructor, as needed, while organizing the course (particularly in stage one). Additionally, NTK could assist the instructor in purchasing/locating course materials and placing these on reserve for student use—this model is typical in other countries and provides an alternative to production of *skripta*.

Funding high-quality instructors would be primary expense for this program. To attract talent, salaries should be competitive. Per half-time instructor (Minimal qualification: Masters degree, preferably in Science or Engineering with 5+ years professional &/or teaching experience): At least Ministry of Education table 2.16.08.12 LEKTOR-INSTRUKTOR (highest level in this salary band). NTK would provide extensive training for any instructors hired as well as provide eVersions of classroom materials and provide open eCourses over time.

- First phase: 0,5 FTE; annual salary plus benefits 480 000 CZK (estimating 40 000 CZK/month for both salary and benefits); materials & miscellaneous expenses: 20 000 CZK; if needed: one-time relocation benefit 50 000 CZK (2 000 EUR); **total minimally 500 000 CZK/annum** *(plus possible relocation expense; then: 550 000 CZK)*

- Second phase: 2x0,5 FTE at the same cost as above; **1 000 000 CZK per annum** *(plus possible relocation expense)*.