Facilitating the discovery of open government datasets through an exploratory data search interface

Monica Swamiraj
monica.swamiraj@alumni.ubc.ca

Dr. Luanne Freund
luanne.freund@ubc.ca

iSchool, The University of British Columbia
470 – 1961 East Mall, Vancouver, BC Canada V6T 1Z1

2015 Open Data Research Symposium, 27th May 2015, Ottawa, Canada

Abstract

The primary area of investigation for this paper is the process of open data discovery, specifically, how novice users search for open government datasets and how this process can be improved. The problem of search on open data portals has featured consistently in Canadian Open Government consultations. A literature review of open data initiatives and processes reveals that open data search is a browsing or investigative task rather than a factual lookup or subject search task. Researchers have proposed exploratory search interfaces as alternatives to traditional search interfaces to support learning and investigative tasks in the information retrieval domain. This paper elaborates on a study conducted to evaluate the usefulness of a visualization-based exploratory data search interface to help novice users discover relevant datasets.

A special feature of this interface is the use of variable names in addition to dataset description for search. Today, data search systems on open data portals tend to rely on the text contained in metadata and dataset descriptions to facilitate keyword search. However, the variable names contain important information about the content and structure of a file. This study also probed the role that variable names could play in search, specifically as a means of facilitating exploration and relevance assessment.

1. Introduction

The Government of Canada (2012) defines Open Government Data (OGD) as "government data that is offered in useful formats to enable citizens, the private sector, and non-government organizations to leverage in innovative and value-added ways. Open data refers to government information that is factual and usually statistical in nature, e.g. population statistics" (Open Data, para. 1). The federal government as well as provincial and municipal governments across Canada are publishing OGD with the intent to make governments accessible and accountable to citizens while improving government services and encouraging innovative uses of these data. However, the consultation reports indicate that not many of these datasets are utilized (Government of Canada, 2013).

A cursory review of the literature indicates that poor findability of datasets is a critical barrier to OGD adoption. The problem of search on open data portals has featured consistently in the Canadian government consultations. In the 2011/2012 government consultations, over 26% of respondents indicated that the government should "[i]mprove the search engine/search capability/search engine optimization" to make finding information online easier (Government of Canada, 2012). Similarly, improvement in search emerged as a key recommendation from stakeholders during the 2013 consultations (Government of Canada, 2013). Searching for OGD is more challenging than searching for other types of content for a number of reasons. As a new form of knowledge dissemination, OGD is unfamiliar to most of its potential users, who may have difficulty conceptualizing what they

are looking for and formulating a specific query. The availability and format of datasets is inconsistent, even erratic, across different portals, leading to more challenges in articulating search parameters and the need to search opportunistically: making use of what exists, rather than finding what one needs. Most importantly, data search is inherently limited by the relative paucity of features that form the basis for most web-based search systems: textual content, hyperlinks, traffic/usage data, etc. As a result, OGD searchers are more likely to browse through lists of datasets to find interesting ones that they can examine further, than to start with a problem or need and seek data on that basis (Government of Canada, 2013). This makes searching for datasets primarily a browsing or investigative task rather than a factual lookup or subject search task.

In the information retrieval domain, researchers have proposed exploratory search interfaces as alternatives to traditional search interfaces to support learning and investigative tasks (Hearst, 2006; Marchionini, 2006; White & Roth, 2009; Wilson, Kules, shraefel and Shneiderman, 2010). Many researchers have also suggested using information visualization techniques to display search results in order to help users both make the connections between search query terms and retrieved documents as well as make decisions pertaining to relevance of those documents (Clarkson, Desai & Foley, 2009; Dörk, Carpendale & Williamson, 2012; Hearst, 2009; Kules & Shneiderman, 2003; Perer & Shneiderman, 2008; White & Roth, 2009) . In this research, we set out to determine if a visualization-based exploratory search interface could be similarly useful for data discovery by novice users, who do not have in-depth knowledge of government datasets.

To this end, we created an exploratory data search interface for numerical datasets that uses a network-based visualization to display datasets and variables matching the search query provided by users. In this context, a variable is a column in the dataset and a variable name is the title of the column. To expose more of the content to discovery, the variable names were indexed along with other dataset metadata such as title, description, publishing organization, etc. In order to make the variable names more accessible to searchers, the variable names rather than dataset names were designated as entry points i.e., each search result corresponded to a variable name rather than a dataset.

The rest of the paper elaborates on the user study conducted to evaluate the efficacy of this strategy and is laid out as follows. First, we review the related work to define the problem of search in OGD portals, understand the exploratory search paradigm, and consider the application of exploratory search to the OGD search problem. Next, we describe the research design of the user study to answer our research question. Finally, we present the findings and provide some recommendations to improve data search.

2. Related Work

2.1. Users, uses and challenges in OGD

The Canadian government has been conducting a series of consultations each year: when developing Canada's Action Plan on Open Government in late 2011 and 2012, before redesigning Canada's open data portal in mid-2013, seeking feedback on year-1 performance in implementing the Action Plan in late 2013, and more recently, when developing the next version of the Action Plan in 2014 (Government of Canada, 2014). We use the reports published by the government compiling the responses received for some of these consultations as key sources in this study to understand who the users of OGD are, in the Canadian context, and what they seek from OGD.

Students, seniors, programmers, public servants, professionals, librarians, civic technologists, independent scholars, several PhDs and MDs, companies making business proposals, and citizens responded to the government consultation in 2011/2012

(Government of Canada, 2012). Of participants, 13% were 25 and under, 31.5% were between 26 and 35, 25.5% were between 36 and 45, 18% were between 46 and 55, 12% were over 56. Their intentions to use OGD were summarized as follows (Government of Canada, 2012):

- In libraries and classrooms for scholastic endeavors,
- In academia to complete their research projects or collaborate with other users (including private sector) in research
- In work related contexts such as for development and marketing related activities, to support recruitment activities, engineering and health care activities
- In a civic participation context, to better understand current affairs, government operations, policies and make informed decisions on subjects such as finances and environment
- In a personal context, related to their interests or hobbies such as travel, culture, heritage, genealogy

Other researchers studying the OGD phenomenon report similar findings. For example, in his frequently cited report on an exploratory study conducted in the United Kingdom, Davies (2010) also observes that "the OGD users span a wide range of contexts" (p. 24) and identifies six overlapping motivations for the use of OGD: government focused i.e. interested in improving government transparency and accountability, technology & innovation focused i.e. interested in creating apps, tools and software, reward focused i.e. interested in competing at hackathons and OGD code-fests, digitizing government focused i.e. improving the processes within the government, problem solving i.e. interested in using data at work or home to solve problems, and social or public sector entrepreneurialism i.e. providing services within or to local/provincial/federal governments, offering services based on OGD for commercial purposes. These motivations are similar to the themes identified in the Canadian government consultations.

Over 80 stakeholders from academia, the software development space, non-profit sector, the technology sector, local and provincial government officials attended the series of government consultations that took place in 2013 in Toronto, Edmonton, Vancouver, Ottawa and Montreal ahead of the Open Data portal redesign (Government of Canada, 2013). These consultations were moderated by David Eaves, an Open Government advocate, who comments that "[t]he diversity of roles and goals of the stakeholders who participated in the events reflect the diverse needs and potential goals the open data portal could, or may even need, to serve". For example, users motivated by civic participation were interested in data about government accountability, users motivated by work context such as software developers and business sectors were interested in data that could "provide business intelligence or front line service information", users from academia asked for more data on provenance, metadata and context of the published OGD while government officials were interested in the process of compiling and sharing data across jurisdictions . This is also evident from the list of datasets requested: data on homeless families and individuals systems, land use data, data on changing traffic patterns, health data, environment related data, data on building and construction, immigration data, postal code data, etc. Eaves notes in his report that many Canadians would "come to the portal with little sense of purpose or potentially experience in working with data".

Across all these user groups, it is clear that finding OGD is a common concern. For example, in the 2011/2012 government consultations, over 26% of respondents indicated that the government should "[i]mprove the search engine/search capability/search engine optimization" to make finding information online easier (Government of Canada, 2012). Similarly, improvement in search emerged as a key recommendation from stakeholders during the 2013 consultations (Government of Canada, 2013). Eaves concisely describes the problem: "Without effective search, users and citizens will never be able to find what they need or are curious to know more about. Reducing the friction cost around finding data is no

guarantee of broad and wide usage, but high friction costs will guarantee little or no use of government data".

However, there is no simple solution. Marchionini, Haas, Zhang & Elsas (2005) succinctly phrase the search problem: "how best to provide highly codified, statistical data to a large, diverse population with varying levels of numerical literacy… To find and understand statistical information, people need context and a means of manipulation that will reveal the story behind the number." (p. 52). Unfortunately, most OGD datasets do not have surrounding context, and even when sufficient context is available, users cannot easily make the connections between multiple datasets. This is because, unlike web pages that are text-heavy, datasets contain more factual, spatial or numerical information and very little text, quite often, just enough to describe the basic metadata.

The 2011/2012 consultation report summarizes the feedback from users on how search systems could be improved: "Respondents also expressed that datasets should be classified in ways that made sense to the user in order to improve search functionality. They indicated a desire for a 'one-stop shop' that provides all government information and has powerful, federated search engines" (Government of Canada, 2012, Organization, para. 2). The 2013 report mentions a problem that many participants shared: knowing that a dataset exists but not being able to find it "without typing an exact phrase or knowing a key term" (Government of Canada, 2013, Improved Search, para. 3). In addition, participants also asked for features such as dataset recommendations that would enumerate related datasets when users look at a specific dataset, thematic search, tagging and navigation of datasets through user-generated tags, grouping of geographic and longitudinal data to identify trends, ability to download sample data for large datasets, and single portal for search across local, provincial and federal datasets. It is apparent that, similar to the motivations for OGD use, the requirements for search are also wide-ranging. They also suggest that finding OGD is a process of exploration for many users.

The data on the motivations of OGD use compiled by Davies (2010) indicate that few users start with a question, i.e., lookup datasets to solve a problem at hand. Rather, they explore available datasets looking for opportunities to innovate and for potential problems to solve. This is also evident in the 2013 consultation report: "While virtually all stakeholders had visited data.gc.ca at some point, only between 50-70% had ever downloaded a dataset and a smaller number, say 15-20% had actually used a dataset in an analysis or application. Some of this was a result of an inability to find interesting datasets" (Government of Canada, 2013, The Data, para. 2).

Hence, we see three types of problems emerging in the context of OGD search:

- Users looking up specific datasets might not know which search terms to use,
- Users starting with a problem at hand might not be aware of the exact dataset that will solve their problem, and
- Users exploring and collecting multiple datasets to create a new app might not be aware of the relationships between datasets.

These search problems align with concept of exploratory search, which encourages learning and investigation (Marchionini, 2006; White & Roth, 2009; Wildemuth & Freund, 2012). Accordingly, we posit that OGD users might benefit more from an exploratory data search interface than a simple text-based search interface. In the next section, we examine exploratory search in more detail.

2.2. Exploratory Search Paradigm

Marchionini (2006) was among the first to formally conceptualize exploratory search. He distinguishes between three types of search tasks: lookup, learn and investigate: "[l]ookup

tasks return discrete and well-structured objects such as numbers, names, short statements, or specific files of text or other media" (p. 42). Whereas "learning searches involve multiple iterations and return sets of objects that require cognitive processing and interpretation … and often require the information seeker to spend time scanning/viewing, comparing, and making qualitative judgments" (p. 43). Investigation searches, on the other hand, "involve multiple iterations that take place over perhaps very long periods of time and may return results that are critically assessed before being integrated into personal and professional knowledge bases … done to support planning and forecasting, or to transform existing data into new data or knowledge … [or] to discover gaps in knowledge so that new research can begin or dead-end alleys can be avoided" (p. 43). Marchionini classifies both learning and investigation tasks as exploratory search tasks.

White and Roth (2009) provide a more specific definition of exploratory search: "a combination of searching and browsing behavior to navigate through (and to) information that helps [searchers] develop powerful cognitive capabilities and leverage their newly acquired skills to address open-ended, persistent, and multifaceted problems" (p. 10). They argue that satisficing, or the principle of least effort, is not applicable to exploratory searches as users are not likely to stop their search as soon as few relevant "information fragments" are encountered.

Regular search interfaces, on the other hand, provide a search result page with a list of results ordered by relevance or another appropriate algorithm. Users seldom navigate beyond the first few results on the list and the systems are not designed to support browsing. They are optimized for precision i.e. "minimizing the number of possibly irrelevant objects that are retrieved" unlike exploratory search interfaces which favor recall i.e. "maximizing the number of possibly relevant objects that are retrieved" (Marchionini, 2006, p. 43). The latter also tend to use a grouping logic, quite often along with some visualization, when displaying search results. Hence, we argue that an exploratory data search interface is more suited to the OGD context than a regular search interface.

2.3. Interface Design

Jiang and Koshman (2008) enumerate the information architectures commonly used in exploratory search interfaces: hierarchical classification, faceted categorization, clustering and social tagging.

- Hierarchical classification requires organization of the information in a rigid non-overlapping hierarchy based on one or more rules, which is expensive and less flexible.
- Faceted classification arranges information according to their attributes, which though more flexible, has to be created at the very beginning.
- Clustering, which is more dynamic, leverages algorithms to group search results but has the risk of mis-labeling.
- Tagging relies on user input i.e., tags for classification, which though inexpensive and flexible, can lead to creation of too-specific groups that may overlap.

Most exploratory search interfaces use one or a combination of at most two information architectures. For example, Hearst (2006) and her colleagues use a hierarchical faceted categorization that is mostly automated but requires manual intervention. Though effective, such a classification requires largely homogenous information, while open government datasets are heterogeneous: not only with respect to the actual data but also with respect to the available metadata and the storage repository used. Without significant manual intervention, a neat non-overlapping faceted categorization is not possible. As governments publish more and more datasets, scalability becomes an important design requirement. Marchionini et al. (2005) also stress the importance of scalability in the context of government statistical information. They use machine-learning algorithms to classify

documents. Although automated clustering leads to some spurious groups, overall, such an approach is more suitable for OGD grouping. Hence, we propose to use a clustering-based information architecture in this exploratory data search interface.

Hearst (2006) posits that "[i]nformation seekers often express a desire for a user interface that organizes search results into meaningful groups, in order to help make sense of the results, and to help decide what to do next" (p. 59). Many exploratory search interfaces use visualization techniques to organize their search results into meaningful groups. Some of the popular techniques are: treemaps, layered mapping with an option to add or remove layers, network visualization, and user modeling (Hearst, 2009; Wilson et al., 2010). After investigating a range of visualization approaches, we opted for a network visualization, inspired by systems such as SocialAction (Perer and Shneiderman, 2008). A network visualization is optimal when there are a large number of nodes and large number of connections but a small number of node or connection types. Since this is indeed the case with OGD and variables, a network visualization is appropriate for an exploratory data search interface.

3. System Design

For the purposes of this study, we only focused on numerical datasets published on different portals. We created a federated search system with two interfaces: an exploratory data search interface and a baseline search interface for comparative evaluation purposes. The federated search system indexes a subset of federal, provincial, municipal and community data portals across Canada. There are at least 70 data portals hosted on a variety of systems: CKAN, Socrata, Microsoft's Open Government Data Initiative and custom solutions. Although the search system is extensible and can support any number of data portals, for this proof of concept we used only a small set of six data portals hosted on CKAN or Socrata portals, which were easy to scrape. The baseline search interface (A) uses lists to display the results (Fig. 1) while the exploratory data search interface (B) uses a network-based visualization to display search results (Fig. 2).

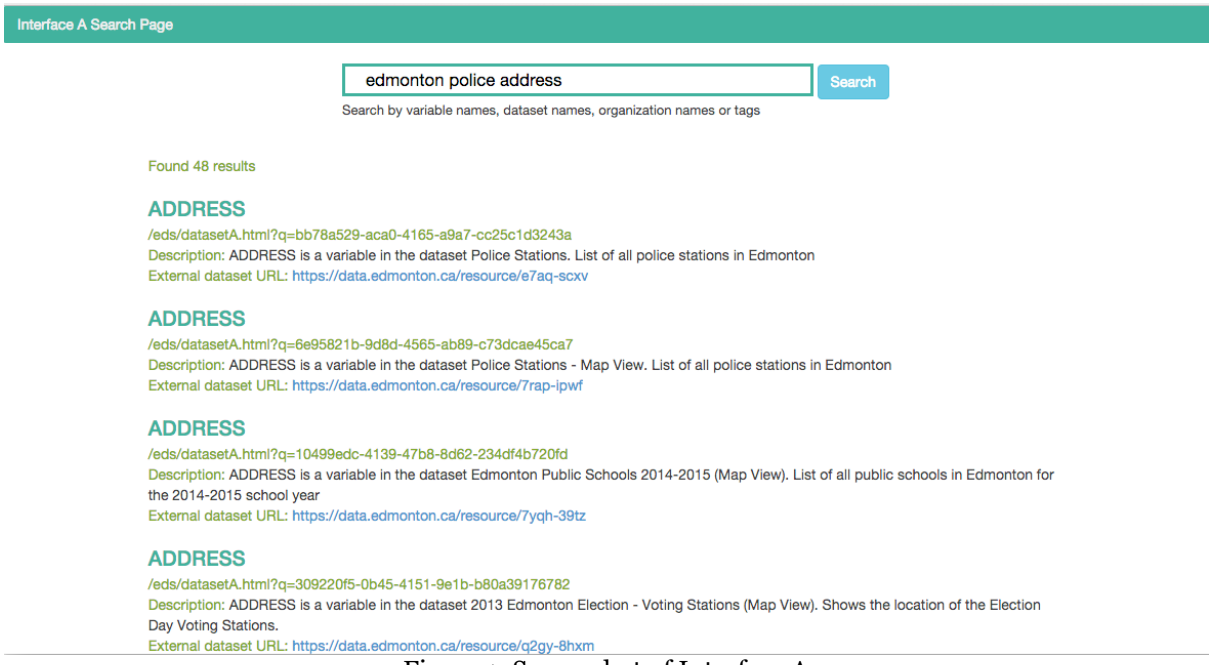

Figure 1: Screenshot of Interface A

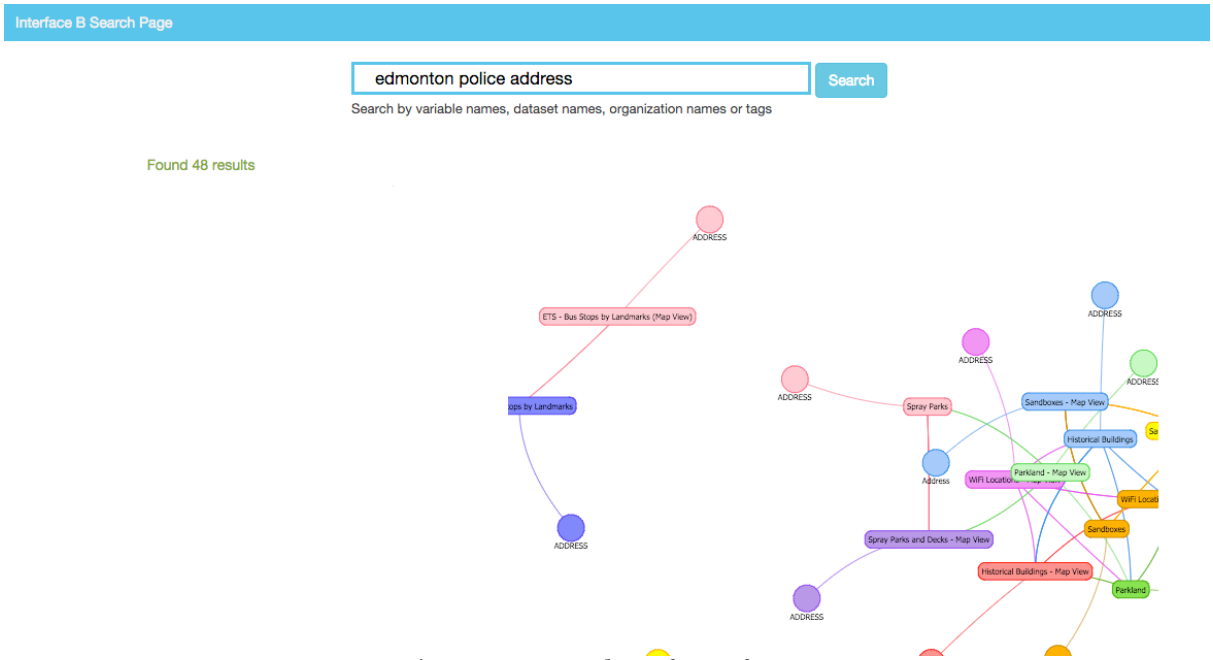

Figure 2: Screenshot of Interface B

Interface B displayed all variable names that matched the search parameters as circular nodes connected with the corresponding datasets, which were displayed as rectangular nodes. All dataset nodes were interconnected based on shared tag information. In other words, two or more datasets that shared more than one tag formed a tightly connected network while datasets sharing no tag information were disjoint. The network-based visualization was zoomable, according to Shneiderman's information visualization mantra: "Overview first, zoom and filter, then details on demand" (Perer & Shneiderman, 2008, p.266). In addition, Interface B allowed users to filter search results according to the name of the publishing organization.

On both interfaces, clicking on a search result displayed relevant information about the dataset such as the title, the name of the publishing organization, associated tag information, links to download dataset file, and a list of all variables in the dataset.

Both interfaces work with the same backend search system that indexed datasets from the Canadian federal open data portal as well as select municipal portals. A stand-alone batch application (Fig. 3) periodically scraped the OGD portals by invoking the CKAN and Socrata APIs to query datasets and the corresponding metadata. Then, it parsed the metadata to identify numerical datasets, downloaded the dataset, extracted the variable names, and indexed each variable name in the search engine along with the corresponding dataset metadata.

The system consisted of three tiers (Fig. 3): a search engine, application logic and a front end. An open source search engine (Elasticsearch) was used to re-index the datasets from different OGD portals in one location. The search engine indexed variable names along with variable and dataset metadata, which enabled search based on both the variable name as well as dataset metadata. The application logic used the search engine APIs to query for variable names that matched the search parameters, and send the query response to the interface. The user interface displayed the search results with variable names acting as entry points but containing other dataset level information such as the title of the dataset, the name of the publishing organization, and description underneath to help users assess the relevance of both the variable as well as the dataset.

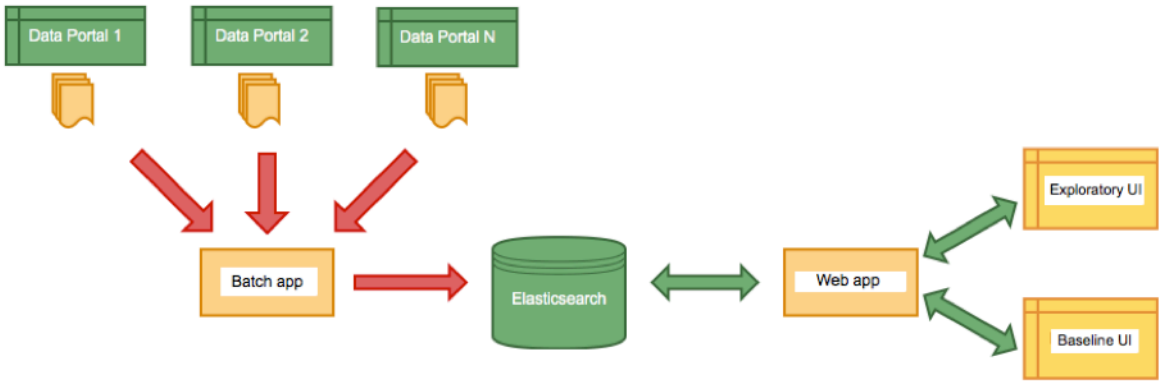

Figure 3: System Architecture of the Federated Search System

4. Method

The primary area of investigation for this study was the process of open data discovery, specifically to determine if an exploratory data search interface is more appropriate than a regular search interface for OGD users who are looking to browse datasets and/or are unfamiliar with the datasets published by federal, provincial and municipal governments. Consequently, our research question was:

> *For novice open data users, how does an exploratory data search interface that visualizes the relationships between datasets perform in comparison with a simple textual search interface?*

In order to answer this research question, we selected an experimental design approach where participants were provided with a set of exploratory data search tasks and asked to complete the search using two interfaces: a baseline interface (A) that displayed a list of datasets as the search result and an exploratory data search interface (B), which visualized the relationships between variable names in the datasets in addition to displaying a list of datasets. As both interfaces used the same underlying search system, the independent variable in this experiment was the interface type. For the experiment, we chose a within-subjects approach, where each participant carried out tasks using both interfaces, to enable comparison between the two interfaces (Kelly, 2009). A secondary variable in the experiment was the task type. We designed four exploratory data search tasks of increasing difficulty level based on the guidelines provided in the literature by Kules and Capra (2009), and Wildemuth and Freund (2012): T1 was the easiest, and T2, T3 and T4 ranged upwards in difficulty (Table 1).

Table 1

Assigned Exploratory Search Tasks

T1	Imagine that you have just moved to Surrey. When moving, you found it hard to choose a neighborhood to live in. You had to look at many different factors that make a neighborhood a good place to live. In order to simplify this process for other people, you are working with your new friends to create a neighborhood quality app based on factors that are important to you, such as safety, transit coverage, availability of bike routes, proximity to farmers markets etc., using the open datasets published by City of Surrey. Please use search interface (A) to find 3-4 relevant datasets.
T2	Imagine that you have been asked to write an article on the issues of immigration in

	Edmonton and the recent federal restriction on immigration of unskilled temporary workers. You are looking for any data available on immigration. You have several ideas to start with such as the size of the labor force, how many people have immigrated in the last 15 years, how many people are currently unemployed, how Edmonton compares with other cities, etc. Please use search interface (A) to find 3-4 relevant datasets.
T3	Imagine that you and your friends are planning to participate in an open data competition. Your team has decided to create an app that visualizes spending in the various Canadian federal government agencies. You are looking for a wide range of government spending datasets related to education, international assistance, etc. Please use search interface (B) to find 3-4 relevant datasets.
T4	Imagine that you have been reading about a discussion on a pharmacare program, to provide financial support for low-income individuals in need of prescription medicine. You would like to find out more about how people across Canada use healthcare including medicine and alternative medicine and their attitudes towards healthcare. Please use search interface (B) to find 3-4 relevant datasets.

We used a Graeco-Latin square design to rotate both interface and task types in order to control for order effects and participant fatigue. In this design, participant P1 first used Interface 1 (baseline) to execute tasks 1 and 2, followed by Interface 2 (exploratory data search interface) to execute tasks 3 and 4 whereas participant P2 first used Interface 1 to execute tasks 2 and 3, followed by Interface 2 to execute tasks 4 and 1, and so on. The task sequences were assigned systematically to participants so that each task was repeated an equal number of times under the two conditions.

4.1. Data Collection

We collected data pertaining to three indicators to answer our research question: participants' familiarity with data search, their ability to complete the exploratory search task, and their subjective feedback on the data search experience. The first indicator allowed us to determine whether the participant is a novice user or an expert user of data, whether the participant has used OGD before and whether the participant is comfortable understanding and interpreting numerical data. The second indicator enabled measurement of the effectiveness of the interface: whether the participant was able to evaluate and choose datasets, and whether the participant faced any functional or usability problems when looking for datasets. For every exploratory search task, a set of datasets were compiled as the solution set, and during the data analysis phase, the datasets chosen by the participants were compared against the solution set to determine completeness. The third indicator, subjective feedback, facilitated the measurement of user satisfaction with the search interface.

We used a combination of observation and questionnaires to collect the necessary data. Screen-capture software was used to record participant interactions with the interfaces. Observation was carried out at playback time using these screen recordings along with the system log data (Kelly, 2009). Questionnaires containing both close- and open-ended questions were used to collect demographic information on participants: their familiarity with data search, their exposure to OGD, their educational qualifications, etc., and also to collect subjective feedback on their interactions with the two interfaces. We used three questionnaires: demographic, post-task and exit questionnaires, to collect the necessary data. The mixed-methods approach allowed us to not only analyze the data for trends but also better understand participants' thought processes.

4.2. Participants

We recruited 16 participants from a graduate program in library, archival and information studies through convenience sampling. Participants were 12 women, 3 men and 1 individual who did not report gender. They were primarily between the ages of 18-29, with 5 in the range of 18-29 and 2 in the range of 40-49. All but one participant had searched for government information previously: 9 participants searched for government information a few times a year, 5 participants searched monthly while 1 participant searched weekly. Participant responses indicated that they would normally search for this type of information using Google and/or search engines provided by government websites. Two participants suggested that they would consult library resources in addition to government websites.

While all participants had previously searched for numerical datasets for school, work or personal interests, only 12 participants indicated that they were comfortable (rating 4 or more on a 7-point scale) reading and understanding numerical or statistical data. Similarly, most participants had previously searched for spatial and temporal datasets were comfortable (rating 4 or more on a 7-point scale) reading and understanding spatial or location-based data and temporal or time-series data. All participants indicated that they were familiar with the concept of open government data and 13 indicated that they had previously searched for open datasets. In comparison to the general population, this was a highly educated study population with strong information and search skills, although non-experts with respect to OGD.

4.3 Protocol

All participants were comfortable with web search; however, none were heavy users of datasets or open data. Each face to face individual session in the study lasted approximately 90 minutes, where participants spent roughly 60 minutes interacting with the search interfaces and 30 minutes answering questions on paper questionnaires.

At the beginning of the study, participants were welcomed and asked to sign a consent form in accordance with the research ethics protocol. Next, a demographic questionnaire was administered to collect background information on participants, their interaction with data on a day-to-day basis, their familiarity with data search, and any prior experience searching for OGD. Third, participants were given a brief introduction to one interface, and then asked to execute two exploratory search tasks with that interface, one at a time. They were given between up to 15 minutes for each task. A screen-capture software was used to record participant interactions with the interfaces. A post-task questionnaire was administered after each task to collect their impressions on their interaction with the search interface. This process was then repeated for the other interface. Finally, after participants completed all tasks, they were asked to provide feedback on their search experience and on the two search interfaces in an exit questionnaire.

4.4. Data Analysis

Participant responses to most questions in the demographic questionnaire were ordinal or categorical data that were transferred to a spreadsheet for analysis. Qualitative responses describing their search processes were open-coded and analyzed for themes (Kelly, 2009). These responses were used to contextualize and interpret the study results. The screen recordings were also open-coded to capture the sequence of steps that participants carried out. These behavioral data map to interaction measures.

The responses to each task were compared against a pre-determined solution set to calculate a level of task completion based on precision and recall metrics. The responses to the post-task questionnaires provided task-level usefulness and usability feedback while the responses to the exit questionnaire provided interface-level usefulness and usability feedback. Similar to the demographic questionnaire, Likert-type questions were coded as

ordinal data and responses to the open-ended questions were open-coded. We analyzed the quantitative data collected through the questionnaires using descriptive data analysis techniques, paired-samples t-tests and one-way ANOVA tests using a significance level of .05. We used Spearman's *rho* to measure correlation between variables. The qualitative data collected through the questionnaires were open-coded and analyzed for themes. The data from the observation, specifically the datasets chosen by participants as being relevant to each task, were compared against a pre-determined solution set to compute precision and recall of the search. The resulting quantitative data was analyzed using descriptive data analysis techniques and ANOVA tests.

5. Results

5.1 User Perceptions

Overall, most participants rated both interfaces favorably with regards to ease of use, learnability and satisfaction with results. On a 7-point Likert scale, 13 participants (80%) gave a rating of 4 or more for ease-of-use of Interface A and Interface B. Similarly, 13 participants (80%) gave a rating of 4 or more for learnability for both interfaces. However, with regards to satisfaction, only 11 participants (68%) gave a rating of 4 or more for Interface A, while 13 participants (80%) gave a 4+ rating for Interface B.

Interestingly, there was a borderline negative correlation between satisfaction for Interface A and Interface B, $r_s(14) = -.49$, $p = .05$, indicating that, while there were no general differences in outcomes by interface, individual participants tended to prefer one interface or the other. This seems to have been influenced by responses to the visualization, as there was a significant correlation between feedback for Interface B and perception of usefulness of the visualization (see Table 2). When asked which interface they preferred, 11 (68.8%) participants mentioned that they preferred Interface B, 4 (25%) preferred Interface A, while 1 (6.3%) participant mentioned that they would prefer B when exploring and A when looking up datasets.

Table 2

Non parametric Bivariate Correlations for Ease-of-Use, Learnability and Satisfaction of Interface B and Perceived Usefulness of Visualization

Variable	1	2	3	4
1. Ease-of-Use B	--	.755**	.915**	.824**
2. Learnability B		--	.849**	.710**
3. Satisfaction B			--	.761**
4. Perceived usefulness of visualization				--

Note: Correlations marked with two asterisks (**) were significant at p < .01 (2-tailed).

Qualitative analysis reinforces the notion that preference was influenced by individual differences in participants. Of the participants who preferred Interface A, one participant commented, "when looking for data I want to get in/out as quick as possible a list is easy to look down & pinpoint something relevant in a glance". Similarly, another participant wrote, "I do not consider myself a visual-based person, so Interface A was easier for me to grasp more quickly". A third participant said, "I wasn't overwhelmed by visuals. I felt more successful and like a better searcher when using A". On the other hand, participants who

preferred Interface B said “[the] presentation of all variables at once [was] useful,” “[it was] easier to see all the variables in a dataset,” “[the interface] gave me a quick intuitive sense of each data set without needing to do much navigation,” “[h]aving the data organized based upon relation to data set made navigation much easier”. These snippets indicate that while some participants found the number of displayed at once in the visualization overwhelming, others found it to be useful, particularly with respect to providing an overview and making associations.

5.2. Completion of exploratory search tasks

Participants were asked to choose 3 to 4 datasets that were relevant to each exploratory search task. However, the total number of relevant results varied by task: T1 had 20 relevant datasets, T2 had 25, T3 had 9, and T4 had 7. In general, participants engaged in exploratory search as evident from an analysis of the screen capture data. For example, they iteratively refined their search queries based on the cues provided in the task descriptions and the metadata for previously found relevant datasets. In addition, participants examined the dataset information page to look up additional information such as names of all the variables, tags etc, and selectively download datasets before making the final relevance decision.

In order to compare the level of completion of search task as well as accuracy of results, we used standard metrics from the Information Retrieval domain: precision and recall, wherein we use Kelly’s (2009) definitions of precision as “[t]he number of relevant retrieved documents divided by the number of retrieved documents” and recall as “[t]he number of retrieved relevant documents divided by the number of relevant documents in corpus” (p. 109).

Overall, the precision scores were relatively high (M= .76, SD= .27, N=62), indicating that about three quarters of the documents selected were relevant. However, the recall scores were low (M= .2, SD= .1, N=62), which is not surprising as we only asked participants to select 3 to 4 datasets. Table 3 displays the means and standard deviations of measures per task and interface.

Table 3

Mean and standard deviation by task and interface (Interface A- baseline; Interface B - exploratory)

Task	T1		T2		T3		T4	
Interface	A	B	A	B	A	B	A	B
Precision	0.86 (0.18)	0.93 (0.15)	0.85 (0.20)	0.81 (0.35)	0.81 (0.21)	0.47 (0.27)	0.76 (0.15)	0.63 (0.32)
Recall	0.17 (0.05)	0.21 (0.09)	0.18 (0.09)	0.13 (0.07)	0.22 (0.12)	0.21 (0.14)	0.27 (0.09)	0.2 (0.11)
Ease of Task	5.13 (1.55)	5.38 (1.19)	4.38 (1.06)	4.88 (1.25)	4.5 (1.41)	4.75 (1.49)	3 (1.31)	3.75 (1.04)
Ease of Interface	4.88 (1.36)	4.88 (1.13)	5.13 (0.99)	5 (1.51)	4.88 (1.46)	5 (1.60)	4 (1.77)	4.5 (1.31)
Satisfaction	5.13 (1.13)	5.13 (1.46)	4.5 (1.31)	4.75 (1.67)	5 (1.51)	5.38 (1.51)	3.88 (0.64)	4.63 (1.9)
n	7	8	7	8	8	8	8	8

To compare the effects of the search interface and task on search outcomes we conducted a one way Anova. There were no significant differences due to the search interface either for

precision (F (1, 60) = 2.85, p = .1) or for recall (F (1, 60) = 1.13, p = .29), indicating that participants were able to accomplish the task at the same level using the two interfaces. Similarly, there were no significant differences in recall due to the task (F (3, 58) = 1.98, p = .13). There was a significant difference due to the task for precision (F (3, 58) = 3.42, p = .02), which seems to be the result of higher precision rates for the easier tasks (T1 M= .9, SD= .16; T2 M= .83, SD= .28) than for the more difficult tasks (T3 M= .64, SD= .29; T4 M= .69, SD= .25). Table 3 indicates that this trend is consistent across both interfaces, although precision seems to drop off more dramatically in the difficult tasks for Interface B than for Interface A.

5.3. Subjective feedback on data search experience

From the feedback provided in post-task questionnaires, on average, the ratings for ease-of-use on a 7 point scale (M=4.78, SD=1.37, N=64) and satisfaction with search results (M=4.8, SD=1.34, N=64) were in the mid-range. A one-way ANOVA to compare the effect of task and search interface on ease-of-use of interface and satisfaction with search results showed no significant differences. This is in line with the result that search outcomes did not differ between interfaces. However, Table 3 suggests that there may be a trend towards higher mean satisfaction scores with Interface B as compared with A as the tasks become more difficult. This is interesting, as Interface B has comparatively lower precision for difficult tasks, but higher satisfaction, suggesting that for more difficult tasks the ability to explore and examine a wider range of datasets, even if they are less relevant, may be preferred by searchers.

A correlation analysis to identify relationships between user perception measures shows that task is negatively correlated with perceived ease of task, $r_s(14) = .44$, $p < .01$) , confirming our own classification of these tasks.

In examining the user comments on the two Interfaces, some common themes emerged. On the one hand, participants indicated that they liked the simplicity and familiarity of Interface A because information on variables is displayed in the search result itself, the results are displayed as lists, and the links appear obvious. On the other hand, participants indicated that they disliked not seeing variable grouping, not being able to filter by publishing organization, not being able to distinguish between datasets that have been visited and those that have not, problem with keywords and retrieving irrelevant results.

Similarly, for Interface B, participants liked one or more of these features: grouping variable names and datasets using colour and links, ability to filter by publishing organization and having the ability to interact directly with the visualization. However, participants disliked the cluttered layout of network and constant movement of the dynamic network visualization. They also felt overwhelmed by the number of results displayed on the screen, and by the fact that the network only occupied one half of the screen. Participants disliked not being able to click on a dataset node directly to view the dataset information.

With regards to making improvements to Interface A, 2 participants suggested grouping of results by nesting variables within datasets. To Interface B, 8 participants (50%) suggested making changes to the network layout by either spacing datasets or optionally displaying variables or making the nodes move less. 7 participants suggested making changes to the screen layout by making the network take up more space, marking the network and making filter options more visible. 3 participants suggested providing an option to dismiss the search results when they were not relevant, while suggested adding additional information to the nodes of graph. At least 3 participants indicated that the movement of nodes of the network should be minimized.

6. Discussion and Implications

In testing this OGD search system with this set of participants, we were hoping to understand how novice OGD users would carry out exploratory search and what features would facilitate success. While the participants are relatively young, well-educated, data-literate and familiar with online searching, this maps quite well onto the typical OGD user profile (Davies, 2010; Government of Canada, 2013), and so the results are of some value in understanding this population. We found that participants were able to complete their assigned tasks quite well using both Interfaces A and B, but that participants seemed to be drawn to one or the other in terms of overall satisfaction. This may be a function of individual differences, such as cognitive style, as it is well-known that field dependence/independence affects the ability to interpret information visualizations (Yuan, Zhang, Chen & Avery, 2011).

There were clear effects of the level of difficulty of the four assigned search tasks, with respect to participants' perceptions and search outcomes. The more difficult the task, the lower the performance and user satisfaction with the results. There is also some evidence that Interface B better supported user needs as the tasks became more difficult, perhaps because it enabled the kind of exploratory, investigative searching required by more difficult tasks. However, further research is needed to confirm this finding.

As expected, there were correlations between participants' satisfaction with search results and ease of use of the search interface. This may be because participants perceive a system to be easier to use when they experience success, or that easier to use systems actually facilitate better retrieval. Either way, a major challenge in data search is presenting inherently complex materials (datasets) in easy-to-use and intuitive interfaces.

Overall, the mean scores for ease-of-use of and satisfaction with search experience were higher for Interface B than Interface A, although differences were not significant. User perceptions of Interface B (satisfaction and ease of use) were associated with their perceptions of the visualization component. Particular features that participants found helpful were: the grouping of variables and datasets, specifically the choice of color as a visual feature; and the ability to narrow down their search results by the publishing organization. In contrast, participants rated the learnability of Interface A more favorably than Interface B, likely based on the familiar format, akin to Web search engines. Specifically, participants found reading information and locating links easier in the list format. This trade-off between appeal and learnability is a well-known point of comparison between visual and textual search interfaces, and is important to take into account when designing OGD applications.

With regards to the usability of the search system, participants found not being able to use operators such as NOT or wildcards, and not being able to search using synonyms very limiting. They also found retrieval of results only partially matching their search terms somewhat confusing.

Participants strongly felt that the usability of search interface A could be improved by using mechanisms to group variable names with the corresponding datasets, to call out the name of the dataset and to filter by publishing organizations. Not being able to distinguish between previously visited links and new links as well as a bug in the pagination plugin also impacted the usability of Interface A. During the observation, we found that participants, when faced with many search results from the same dataset, tended to skip pages to avoid the 'repetitive' results and in the process missed other relevant datasets.

With regards to the usability of search interface B, participants asked that the view-window of the visualization be broadened so that more nodes would be visible. Participants also

asked that the drop-down to filter by publishing organization be made more visible. Participants found not being able to click directly on a dataset node very frustrating, as this increased the number of clicks to access the dataset. Most participants also tended to click on external dataset URL rather than the title of result to access dataset description. Finally, with regards to the graph, participants asked that the font size be increased and nodes not overlap, to improve the readability of dataset as well as variable names. They requested a feature to dismiss or hide irrelevant results so as to decrease the clutter on screen. Participants also found the constant movement of nodes on screen, especially in the case when a large number of search results were displayed, disconcerting.

Participant feedback indicates that high-level decisions were being made at the dataset level, which could be because the search task required them to choose relevant datasets. However, the variable names and publishing organization also play a role in the relevance decision for datasets. Many participants rated Interface B favorably largely due to the ability to narrow down search results by publishing organization: "A would be preferred with grouping by source or filtering," "[b]ecause of the ability to filter by creator," "[l]ack of filtering made searching very difficult," etc. Participants commented on the cryptic nature of variable names, "heading often times not a title (ex. FROM)", "[n]ames are sometimes very undescriptive (ex. A whole pg listing 2011 as the main entrance point)" and relied on the description for more information on both the variable as well as the dataset.

Finally, many participants were not familiar with the Canadian open datasets. Although this lack of familiarity caused participants to continue to refine their searches and explore the available datasets until they found 3-4 relevant datasets, participants were frustrated when many of their queries did not yield the results that they had expected. For example, for task T3 on Canadian government spending, many of the participants commented that knowing the names of the different departments would have helped them formulate better search queries.

Some design implications for OGD search arise from these results. Since user preferences seem to be influenced by individual differences, it may be best to design combined interfaces with both listings and visualizations, to provide options for users. The challenge is to create a mapping between the search results displayed on the visualization and those displayed in the list. Though the effectiveness of using variable names as entry points was limited by the very general or cryptic nature of many of them, participants did appreciate the groupings of datasets and variable names. Finding more creative and flexible ways to leverage variable names and associated metadata and use it to determine associations between datasets is likely to be valuable. Finally, although participants rated the exploratory search interface more favorably than the list interface, there was no significant impact on the effectiveness and user satisfaction due to the exploratory search interface. This raises the question of the role of such interfaces, which may be best suited to particular domains or disciplines, or to particular user tasks. Rather than opting for a "one size fits all approach," more work to match the tools to the tasks is needed to make OGD findable and usable.

7. Conclusion

In this paper, we presented the problem of OGD search and proposed the use of exploratory search to enable novice users to explore open datasets. In order to evaluate the efficacy of exploratory search, we created a search system and two interfaces: a baseline interface and an exploratory data search interface. The search system allowed users to search for relevant datasets using both metadata such as title, description, publishing organization etc., and variable names. Also, the search system used variable names as entry points i.e., the title of search results were variable names and not dataset names.
Results show that although participants rated the exploratory data search interface more favorably than the list interface, there was no significant impact on the effectiveness and user

satisfaction due to the exploratory search interface. In fact, participants showed a marked preference for one of the two interfaces depending on whether they identified themselves as "visual" or not. Participants did find having access to the variable names within datasets useful when making relevance decisions, but, they were frustrated with variable names as entry points as the tasks specifically required choosing 3-4 datasets. More work needs to be done to refine the representation of this information to provide optimal support for searchers.

This was a small study, with limited generalizability. It focused on exploratory search behavior, which is only one mode of searching. A visual interface is much less likely to be useful for finding specific information or bulk downloading of datasets, for example. The study used convenience sampling to recruit 16 participants, all of whom were graduate students. Hence, the participants are not representative of all OGD users in Canada. The search system itself was fully functional, but it used a small collection and it was not optimized. Participants were frustrated as they could not use some familiar features of search systems, which may have negatively impacted their user satisfaction ratings for both interfaces. There were additional hardware and software issues often associated with prototypes that may also have influenced user perceptions.

Future work might include further assessment of the system after optimizing the visualization as well as the search system, which might eliminate the effect of the less-than-optimal user experience. A combined interface could be designed to see if users would select different results displays depending on their personal preferences or to suit different search tasks or domains. An important goal of this work is to better understand how users make sense of datasets and of the relationships between multiple datasets while searching, so that they can discover new and useful data without already knowing that it exists. As open government datasets are only recently gaining popularity, the various government data portals are still evolving. The presence of many different data storage platforms such as CKAN, Socrata, Microsoft OGDI etc., allows governments to migrate from one platform to another depending on their use cases. However, this also leads to less standardization with respect to metadata and developers APIs. A production-ready search system should factor in this dynamic nature of the open data portals and support querying from a range of data storage platforms.

References

Clarkson, E., Desai, K., & Foley, J. D. (2009). Resultmaps: Visualization for search interfaces. *Visualization and Computer Graphics, IEEE Transactions on, 15*(6), 1057-1064. doi: 10.1109/TVCG.2009.176

Davies, T. (2010). *Open data, democracy and public sector reform. A look at open government data use from data.gov.uk*. Retrieved from http://www.opendataimpacts.net/report/wp-content/uploads/2010/08/How-isopen-government-data-being-used-in-practice.pdf

Dörk, M., Carpendale, S., & Williamson, C. (2012, May). Fluid views: a zoomable search environment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 233-240). ACM. doi: 10.1145/2254556.2254599

Government of Canada. (2012). *Open government consultation report*. Retrieved from http://data.gc.ca/eng/open-government-consultation-report

Government of Canada. (2013). *Open data roundtables summary report*. Retrieved from the Canadian Government data portal: http://data.gc.ca/eng/open-data-roundtables-summary-report

Government of Canada. (2014). *Canada's action plan on open government 2014-16*. Retrieved from http://open.canada.ca/en/content/canadas-action-plan-opengovernment-2014-16

Hearst, M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM, 49*(4), 59-61. doi: 10.1145/1121949.1121983

Hearst, M. (2009). Search user interfaces. Cambridge University Press.

Jiang, T., & Koshman, S. (2008). Exploratory search in different information architectures. *Bulletin of the American Society for Information Science and Technology, 34*(6), 11-13.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1-2), 1-224. doi: 10.1561/1500000012

Kules, B., & Capra, R. (2009, June). Designing exploratory search tasks for user studies of information seeking support systems. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* (pp. 419-420). ACM. doi: 10.1145/1555400.1555492

Kules, B., & Shneiderman, B. (2003, May). Designing a metadata-driven visual information browser for federal statistics. In *Proceedings of the 2003 annual national conference on Digital government research* (pp. 1-6). Digital Government Society of North America.

Marchionini, G., Haas, S. W., Zhang, J., & Elsas, J. (2005). Accessing government statistical information. *Computer, 38*(12), 52-61.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41-46. doi: 10.1145/1121949.1121979

Perer, A., & Shneiderman, B. (2008, April). Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 265-274). ACM. doi: 10.1145/1357054.1357101

White, R. W., & Roth, R. A. (2009). Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *1*(1), 1-98. doi:10.2200/S00174ED1V01Y200901ICR003

Wildemuth, B. M., & Freund, L. (2012, October). Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval* (pp. 1-10). ACM. doi: 10.1145/2391224.2391228

Wilson, M. L., Kules, B., schraefel, m.c., & Shneiderman, B. (2010). From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, *2*(1), 1-97. doi: 10.1561/1800000003

Yuan, X., Zhang, X., Chen, C., & Avery, J. M. (2011). Seeking information with an information visualization system: a study of cognitive styles. *Information Research: An International Electronic Journal*, *16*(4), n4.