EXPLORING LEARNING OPPORTUNITIES IN OPEN DATA USE ACROSS DATA LITERACY LEVELS

by

Ak Wai Li

B.Sc., The Chinese University of Hong Kong, 2008
GradDip.AppSc., Charles Sturt University, 2011
MInfoStud., Charles Sturt University, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF LIBRARY AND INFORMATION STUDIES

in
THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2020

© Ak Wai Li, 2020
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis entitled:

_Exploring learning opportunities in open data use across data literacy levels_

submitted by Ak Wai Li in partial fulfillment of the requirements for
the degree of Master of Library and Information Studies
in Library and Information Studies

Examing Committee:

Dr. Luanne Sinnamon, School of Information, UBC
Supervisor

Dr. Richard Kopak, School of Information, UBC
Supervisory Committee Member
Abstract

The purpose of this study is to explore learning opportunities for users with different data literacy levels when using open data within an open data portal, with the goal of informing design. An open data portal is a web-based data repository provided by a government to allow the public to find and use open-access government data. Three aspects are investigated: user interactions with the open data portal and associated tools; obstacles faced in interacting with the portal; and learning strategies employed. The City of Vancouver’s open data portal was the study environment. Fourteen university students participated in this remotely administered study which included an assigned task, a think-aloud observational session, a self-completion questionnaire, and a semi-structured interview. There were divvied into high/low data literacy groups based on a self-assessment questionnaire. The video recordings of the think-aloud observational sessions and semi-structured interviews were analyzed using qualitative methods. Participants in the low data literacy group faced more obstacles due to gaps in understanding how to process open data than those in the higher data literacy group. Participants in the high data literacy group mainly faced obstacles in figuring out how to use the unfamiliar data analysis tools in the portal. Some important data processing activities such as evaluating data quality, data wrangling, labeling analyzed findings and data citation, seemed to be lacking across all participants, regardless of their data literacy self-assessment. These results suggest that an open data portal should improve usability to help users learn to use the tools and provide learning tools to help users develop needed data literacy skills to make good use of open data. The findings have implications for the future study of designing learning tools in an open data portal for users with different data literacy levels.
Lay Summary

This study explores opportunities for including learning tools in open data portals to support users with different levels of data literacy. The study investigates if there are any differences in user interactions with an open data portal and other systems, obstacles faced, and ways to overcome obstacles among people with different levels of data skills when they use open government data to answer an enquiry related to employees’ salary. The study finds that users with low data skills need external support to learn to use open government data properly. The findings help to generate some possible learning tools to be designed in an open data portal.
Preface

This thesis is original, unpublished, independent work by the author, Ak Wai Li. Dr. Luanne Sinnamon and Dr. Richard Kopak provided recommendations for the design of the study and data analysis. The research was approved by the University of British Columbia Behavioral Research Ethics Board under the project title “Using Open Government Data (Online Study)”, with the certificate number H20-01102.
# Table of Contents

Abstract .......................................................................................................................... iii
Lay Summary ................................................................................................................ iv
Preface ............................................................................................................................ v
Table of Contents .......................................................................................................... vi
List of Tables .................................................................................................................. x
List of Figures ................................................................................................................. xi
List of Abbreviations ..................................................................................................... xii
Glossary ......................................................................................................................... xiii
Acknowledgements ....................................................................................................... xiv

Chapter 1: Introduction .................................................................................................. 1
  1.1 Background .............................................................................................................. 1
  1.2 Significance of this research .................................................................................. 3
  1.3 Goal of the research ............................................................................................... 4

Chapter 2: Literature review ......................................................................................... 5
  2.1 Data literacy competency framework ...................................................................... 5
  2.2 Data literacy skill assessment ................................................................................ 7
  2.3 Design of open data portals .................................................................................. 9
  2.4 Informal learning ................................................................................................... 12
  2.5 Research questions ............................................................................................... 15

Chapter 3: Research Methods ....................................................................................... 17
  3.1 Overview of research design ................................................................................ 17
3.2 Selection of an open data portal .................................................................................. 18
3.3 Task and dataset selection .......................................................................................... 22
3.4 Study procedures and instruments ............................................................................. 25
  3.4.1 Think-aloud observational session ........................................................................ 25
  3.4.2 Semi-structured interview ................................................................................... 26
  3.4.3 Study procedures .................................................................................................. 27
3.5 Recruitment of participants ....................................................................................... 27
  3.5.1 Classifying participants based on their data literacy levels .................................. 28
3.6 Data Analysis ............................................................................................................. 30
  3.6.1 Processing Camtasia recordings .......................................................................... 30
  3.6.2 Analyzing entries ................................................................................................. 32
  3.6.3 Processing self-completion survey ....................................................................... 34

Chapter 4: Research findings ............................................................................................ 37
4.1 Participants’ interactions .............................................................................................. 37
  4.1.1 Participants’ interactions to find data .................................................................. 37
  4.1.2 Participants’ interactions to use data ................................................................... 39
  4.1.3 Summary of participants’ interactions ................................................................ 42
4.2 Obstacles faced ........................................................................................................... 43
  4.2.1 Finding a dataset .................................................................................................. 43
    4.2.1.1 Where to search .............................................................................................. 43
    4.2.1.2 How searching works .................................................................................... 45
    4.2.1.3 Selecting the most appropriate dataset .......................................................... 47
  4.2.2 Using the dataset ................................................................................................ 49
4.2.2.1 Wrangling data ................................................................. 50
4.2.2.2 Analyzing data ................................................................. 52
4.2.2.3 Presenting analyzed findings ............................................ 57
4.2.3 Summary of obstacles faced .............................................. 59
4.3 Learning strategies .............................................................. 61
  4.3.1 Trial-and-error ................................................................. 61
  4.3.2 Step-by-step instructions .................................................. 63
  4.3.3 Asking for advice ............................................................ 66
4.4 Summary of research findings .............................................. 67

Chapter 5: Discussion ................................................................. 70
  5.1 Key insights of research findings .......................................... 70
    5.1.1 Profile of low and high data literacy users ....................... 70
    5.1.2 Obstacles ................................................................. 72
    5.1.3 Learning Strategies .................................................... 75
    5.1.4 Opportunities for supporting learning ........................... 77
  5.2 Challenges and limitations ................................................ 84

Chapter 6: Conclusion ............................................................... 86

Bibliography .............................................................................. 89

Appendices .................................................................................. 98
  Appendix A User study documents ........................................... 98
    A.1 Study instructions of think-aloud observational session ........ 98
    A.2 Self-completion questionnaire .......................................... 101
    A.3 Researcher’s script during the study session ....................... 105
Appendix B Codebooks .................................................................................................................. 109

B.1 Codebook 1 for coding tools used to find data ............................................................... 109

B.2 Codebook 2 for coding tools used to use data .............................................................. 110

B.3 Codebook 3 for coding obstacles that participants faced ............................................. 111

Appendix C Findings of the study .............................................................................................. 112

C.1 Demographics of participants ......................................................................................... 112

C.2 Self-assessment data literacy survey among each participant ...................................... 113

C.3 Datasets visited by participants ..................................................................................... 114

C.4 External websites visited by participants ..................................................................... 115
List of Tables

Table 3.1 Evaluation of usability of CoV’s open data portal. .......................................................... 19
Table 3.2 Evaluation of user engagement of CoV’s open data portal. .......................................... 20
Table 3.3 Participant N01’s extract of transcript entries of the think-aloud session ................. 31
Table 3.4 Participant N01’s extract of transcript entries of a semi-structured interview .......... 31
Table 3.5 Demographics of participants compared with their self-assessed data literacy groups 35
Table 4.1 Participants’ interactions with CoV’s open data portal to find ERE-75K dataset........ 38
Table 4.2 Summary of search queries among the order of attempts............................................ 46
List of Figures

Figure 2.1 Kolb’s experiential learning cycle. Adapted from Kolb, 1984. ........................................ 14
Figure 2.2 Information search process. Adapted from Vakkari, 2016, p. 9................................. 14
Figure 3.1 Data literacy assessment................................................................. 29
Figure 4.1 Webpages visited and time spent to find the ERE-75K dataset among participants... 39
Figure 4.2 Data analysis tools used among participants .............................................. 40
Figure 4.3 Number of search queries among participants .............................................. 45
Figure 4.4 Participant N11’s graph for ‘Planner Ii’ in the write up ................................. 50
Figure 4.5 Participant N11’s graph for ‘Planner II’ in the write up ................................. 51
Figure 4.6 Participant N10’s table for Firefighter salaries in the write up ....................... 51
Figure 4.7 Participant N11’s stacked line chart and the write up ...................................... 53
Figure 4.8 Participant N11’s normal line chart........................................................ 53
Figure 4.9 Participant N02’s graph by job titles ........................................................ 54
Figure 4.10 Participant N02’s graph by departments .................................................... 55
Figure 4.11 Participant N07’s graph ......................................................................... 56
Figure 4.12 Participant N09’s graph ......................................................................... 57
Figure 4.13 Participant N13’s write up ....................................................................... 58
Figure 4.14 Participant N08’s write up ....................................................................... 59
Figure 5.1 Kolb’s learning cycle in the context of using the ‘Analyze’ tool in the CoV’s open
data portal. Adapted from Kolb, 1984. ........................................................................ 79
**List of Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>CBE</td>
<td>Council budget and expenses</td>
</tr>
<tr>
<td>CoV</td>
<td>City of Vancouver</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-separated values</td>
</tr>
<tr>
<td>ERE-75K</td>
<td>Employee remuneration and expenses, earning over $75,000</td>
</tr>
<tr>
<td>FAQ</td>
<td>Frequently asked questions</td>
</tr>
<tr>
<td>OGD</td>
<td>Open government data</td>
</tr>
<tr>
<td>SA-HDLG</td>
<td>Self-assessed higher data literacy group</td>
</tr>
<tr>
<td>SA-LDLG</td>
<td>Self-assessed lower data literacy group</td>
</tr>
<tr>
<td>VPL</td>
<td>Vancouver Public Library</td>
</tr>
<tr>
<td>VPLB</td>
<td>Vancouver Public Library Board</td>
</tr>
<tr>
<td>WPRG</td>
<td>Workforce pay rates and gender</td>
</tr>
</tbody>
</table>
**Glossary**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data literacy</td>
<td>An individual’s skill set to understand, find, obtain, read, interpret,</td>
</tr>
<tr>
<td></td>
<td>evaluate, manage, and use data.</td>
</tr>
<tr>
<td>Design thinking</td>
<td>A systematic and collaborative approach to identifying and creatively</td>
</tr>
<tr>
<td></td>
<td>solving problems.</td>
</tr>
<tr>
<td>Microdata</td>
<td>Non-aggregated data of a dataset</td>
</tr>
<tr>
<td>Open data</td>
<td>Digital data that is made available with the technical and legal</td>
</tr>
<tr>
<td></td>
<td>characteristics necessary for it to be freely used, reused, and</td>
</tr>
<tr>
<td></td>
<td>redistributed by anyone, anytime, anywhere.</td>
</tr>
<tr>
<td>Open data portal</td>
<td>A web-based data repository to store and provide access to open data.</td>
</tr>
</tbody>
</table>
Acknowledgements

I offer my enduring gratitude to my supervisor, Dr. Luanne Sinnamon, for her dedication, support, and guidance. I also thank my supervisory committee member, Dr. Richard Kopak for his constructive feedback.

I am grateful to the UBC’s Designing for People (DFP) Research Cluster for offering me fellowship, which helped me to overcome financial hardship during the COVID-19 pandemic.

Special thanks are owed to my husband, my daughter, my family who have supported me throughout my years of education.
Chapter 1: Introduction

1.1 Background

Open data is a global trend to increase government transparency. With the advancement of digital technology and the implementation of e-government, a great variety of government data, such as environmental, housing, education, and economic data, is created in digital formats in the process of day-to-day government operations. To maximize the reuse of data and generate greater value, the open data movement aims at opening data to the public with no restrictions (Attard, Orlandi, Scerri, & Auer, 2015). Open data refers to “digital data that is made available with the technical and legal characteristics necessary for it to be freely used, reused, and redistributed by anyone, anytime, anywhere (Open Data Charter, 2015). In 2009, some governments initiated an open government and open data strategy to strengthen democracy, and to improve governments’ efficiency and effectiveness (Executive Office of the President of The United States, 2009; Government of Canada, n.d.; Her Majesty Government, 2009; The Government 2.0 Taskforce, 2009). After a decade of the open data movement, 30 governments all around the world such as Canada, UK, Australia, and Japan are open data adopters (World Wide Web Foundation, 2018). Open data adopters and advocates believe that open data can generate political and social benefits, economic growth, and optimize operational and technical processes (Janssen, Charalabidis, & Zuiderwijk, 2012). The principles of open data include that data should be (i) open by default, (ii) timely and comprehensive, (iii) accessible and usable; (iv) comparable and interoperable; (v) for improved governance and citizen engagement; and (vi) for inclusive development and innovation (Open Data Charter, 2015).

Despite the claim, open data has not yet met the basic requirements to generate values and benefits after a decade of the open data movement (Open Data Barometer, 2017) as one of the
major challenges is that many citizens outside government organizations do not have sufficient data literacy skills to make use of open data (Janssen et al., 2012). To transform open data into positive value, data users should have sufficient data literacy skills, which enable individuals to “access, interpret, critically assess, manage, handle and ethically use data” (Calzada Prado & Marzal, 2013, p. 126). Data users also require certain hardware and software to download, process, and analyze open data. Although the open data movement aims to engage everyone for civic participation, people with low data literacy skills and/or limited access to digital infrastructure are excluded from using open data (Gurstein, 2011; Open Data Barometer, 2017). If there are no corresponding measures to support these users, only privileged individuals and enterprises that have the necessary resources and skills can get benefits from open data (Johnson, 2014). To avoid creating a data-divide and inequalities, governments should provide access to data processing tools (Ubaldi, 2013), guidance (Janssen et al., 2012), and data literacy training (World Wide Web Foundation, 2018) to help inexperienced data users make effective use of open data.

To promote data literacy among citizens, some governments have partnered with universities, libraries, schools, and other institutions/organizations/groups within civil society to create data literacy courses (Gascó-Hernández, Martin, Reggi, Pyo, & Luna-Reyes, 2018; Humphrey, 2005; Ridsdale et al., 2015). Some governments organize events such as conferences, workshops, and hackathons to engage and teach citizens to use open data (Connor & Wiley, 2018; Government of British Columbia, 2018). To acquire new or develop existing skills and knowledge, learners should engage in an on-going learning process. A “one-shot” instruction is not sufficient to develop data literacy skills as learners need repeated exposure to data literacy concepts (Wanner, 2015). Further, it is unclear that learners will continue to develop and use new skills after the completion of formal training (Stephenson & Schifter Caravello, 2007). Thus, the effectiveness
of face-to-face instruction is questionable. Moreover, only a small proportion of learners can be reached through face-to-face instruction as such sessions are offered at limited times and at designated locations to limited groups of participants who have time and/or money to attend events.

1.2 Significance of this research

Open data portals have the potential to support users to develop data literacy skills informally when the users are making use of open data within the open data portals. Open data portals primarily serve as a gateway for the public to search for open data. Some open data portals such as City of Vancouver’s Open Data Portal (City of Vancouver, n.d.), and City of Edmonton’s Open Data Portal (City of Edmonton, n.d.) also provide built-in tools to analyze data and offer video guides or online courses to learn about open data use. In information seeking, people perform different types of searches such as fact-finding, specified search, and unfocused search (Rieh, Collins-Thompson, Hansen, & Lee, 2016), in which the searching activities may trigger various learning behaviours ranging from lower level cognitive activities, such as remembering and understanding to higher level ones, such as analyzing and creating (Bloom & Krathwohl, 1984). They are accomplishing tasks and informally learning throughout the searching activity. Considering that informal learning is embedded in searching activities, Rieh et. al. (2016, p. 28) suggested that “search systems could be transformed as a learning technology that could support critical and creative learning through various kinds of search behaviour. It is suggested that the e-government domain should move beyond search to information interaction and use, and to provide an unobtrusive learning environment for the user (Freund, Kopak, Kessler, & Vantyghem, 2013). However, a study found that some open data portals lack support features to help users to learn to use open data (Máchová, Hub, & Lnenicka, 2018). If open data portals can be designed to support
users to informally learn to use data, users can build up data literacy skills at anytime and anywhere when they interact with the portals.

1.3 Goal of the research

The main goal of this research is to discover learning opportunities in open data use across users with a range of data literacy levels. In order to design a user-centered data portal that supports the development of data literacy skills through the use of the system and data, it is necessary to first identify how users interact with the system and data and what obstacles users face. Empathizing how users think and feel helps to define challenges to be addressed so that system designers can focus on generating ideas to create solutions to the defined challenges (Institute of Design at Stanford, n.d.). Through discovering how users with different data literacy skills make use of open data, this research will provide insights for the development of learning tools within the open data portals to support users to develop data literacy skills through the use of open data.
Chapter 2: Literature review

In this chapter, I review prior research on data literacy and the design of open data portals. Open data portals are primarily search systems to retrieve open data. To explore the possibility of transforming open data search systems into open data search and learning systems, it is important to identify core skills that users should have to make proper use of open data. Thus, in section 2.1, I first introduce the definition of data literacy and its competency framework. In section 2.2, I review existing data literacy assessment tools to measure an individual’s data literacy skills. In section 2.3, I review the evaluation criteria for a useful open data portal that was used in the selection of an open data portal for this study. In section 2.4, I review the informal learning model to understand how people learn without formal instruction.

2.1 Data literacy competency framework

Data literacy is inter-related with statistical literacy and information literacy (Shields, 2005). From the 1990s through the 2000s, data literacy was considered to be equivalent to statistical literacy which is defined as the ability to access, understand, and critically evaluate statistical results and data-related arguments (Gray, 2004; Stephenson & Caravello, 2007; Wallman, 1993). However, statistical literacy is largely concerned with summarized numerical data (Shields, 2005). Although most authors agree that statistical literacy should be a component of data literacy (Calzada Prado & Marzal, 2013; Fontichiaro & Oehrli, 2016; Shields, 2005), equating data literacy to statistical literacy is limited as it excludes other types of data including non-summarized numerical data and qualitative data. Data literacy is a broader concept than statistical literacy as it is about individuals’ abilities to handle a wider range of data, including statistics, non-summarized numerical data, and qualitative data. Data literacy is defined as an individual’s skill set to “access, interpret, critically assess, manage, handle and ethically use data” (Calzada Prado & Marzal, 2013,
The characteristics of data lay the foundations for defining information because organizing, processing, and adding value to data within a given context transforms data into meaningful and useful information to people (Rowley, 2007). Gummer & Mandinach (2015) defined data literacy as the ability to transform data into information for decision making by collecting, understanding, analyzing, interpreting, and ethically using data. Information literacy is a set of integrated abilities to discover, understand, and use information (Association of College and Research Libraries, 2015). Information can be presented in different formats such as text, video, audio, image, and data. Information literacy is a broader concept than data literacy as it is about individuals’ abilities to handle a wider range of information. Statistical literacy is a component of data literacy while data literacy is a component of information literacy. To help users to transform open data into valuable information, the learning tools in the open data portal should focus on developing users’ data literacy skills rather than information literacy or statistical literacy.

The data literacy competency framework has five core competencies, including (1) understanding data, (2) finding and/or obtaining data, (3) reading, interpreting and evaluating data; (4) managing data, and (5) using data (Calzada Prado & Marzal, 2013). In Calzada Prado & Marzal’s data literacy competency framework (2013, p.130-131), a data-literate person needs to:

(i) understand data, including
   a. knowing how to define data and various possible types of data;
   b. knowing how data is generated in society and the implications of data use;

(ii) find and/or obtain data, including
   a. selecting the most relevant data sources from all possible data sources; and
   b. obtaining data from existing data sources or undertaking research.
(iii) read, interpret and evaluate data, including
   a. reading and interpreting various forms of data presentation such as written, numerical or graphic presentation;
   b. evaluating data critically such as authorship and data collection methods;

(iv) manage data, including
   a. knowing the need for data management for subsequent reuse;
   b. managing data with metadata

(v) use data, including
   a. preparing data for analysis by using necessary data analysis tools;
   b. presenting the results of data analysis;
   c. making ethical use of data such as citing data sources.

2.2 Data literacy skill assessment

The data literacy framework exists, but there is a lack of any validated data literacy skills assessment from the literature. Two approaches in assessing data literacy skills are found, including self-assessment and scenario-based assessment.

Self-assessment is a common approach to measure an individual’s data literacy skills as it is straightforward and easily implemented. Data To The People, an Australia-based company, developed a free online self-assessment tool named Databilities in 2018 (Data To The People, 2020). In using Databilities (n.d.), an individual is assessed on three important activities aligned with data literacy: (1) data reading, (2) data writing, and (3) data comprehension. Each section corresponding to these three activities has between 5 to 12 questions. All questions ask the respondent to choose one of several given statements that best describe himself/herself. Take one question from the “data writing activities” as an example:
“Which of these statements best describe you?

☐ With guidance, I can collect simple data in a format provided to me.

☐ I can collect simple data in a format provided to me.

☐ I can collect data in simple and more complex forms.

☐ I can collect data in a variety of forms to support my needs.

☐ I can assist others to collect data in a simple form to support their needs.

☐ I can teach and assist others to collect data in a variety of forms to support their needs.

☐ None of these describe me.” (Data To the People, n.d.)

After answering all the questions, the respondent gets a result indicating their overall data literacy level, reading level, writing level, and comprehension level where 1 indicates that an individual “can complete simple tasks with instructions” and a 6 indicates that an individual “can teach and assist others to complete complex problems and tasks” (Data To the People, n.d., p. 3).

Qlik, a business intelligence software company, offers a free online data literacy test called The Data Literacy Project (Qlik, n.d.-b). The Qlik (n.d.) test contains 10 multiple-choice questions about the respondent’s experience of working with data, their feelings toward data, and their statistical knowledge. After completing the test, the respondents receive a title such as data dreamer and data knight, which are intended to indicate levels of competence.

Although it only takes a few minutes to complete these types of self-assessment tests, self-evaluation of competence often varies from the truth as people may be ignorant of their incompetence and overestimate their actual abilities (Dunning, 2011). Moreover, the validity and reliability of these self-assessment tests are unknown because there is no transparency on how these tests were developed and evaluated by these companies which sell data-related products.
and services such as software, consultations, and training sessions to organizations (Data To The People, 2020; Qlik, n.d.-a).

Scenario-based assessment is an alternative, more objective, approach to measuring an individual’s data literacy skills. For example, the U.S. Department of Education conducted a study to assess data literacy among teachers in the United States. A sample of teachers participated in a 45-minute interview with the first 15 minutes asking teacher’s personal experience with data use and the following 30 minutes asking the teacher to solve two scenario-based problems using the given data table or graph while thinking aloud at the same time (Means, Chen, DeBarger, & Padilla, 2011). The scenarios were designed to evaluate five data literacy components, including finding data, reading data, interpreting data, using data, and creating questions that will generate useful data. Teachers’ responses were analyzed by comparing them with the desired responses such as finding the actual data from a complex graph accurately (Means et al., 2011, p. 78). Means et al. (2011) stated that this study was exploratory, and that future research is needed to evaluate this type of data literacy assessment.

So, while data literacy frameworks exist, there is a lack of validated instruments to assess data literacy among the general population. Self-assessment is a simple, quick, and generic approach, but subject to respondent bias. While more objective scenario or test-based approaches exist, these are time consuming to administer and we did not find any such instruments to be available for use.

2.3 Design of open data portals

Open data portals serve as platforms to connect governments with all citizens. Therefore, it is important to design open data portals for citizens that possess a wide range of skills and abilities. Many governments set up open data portals as searchable catalogues with metadata
about datasets and dataset download options to allow citizens and stakeholders to easily access open data online (Máchová et al., 2018). Open data portals have become the primary access point for citizens to use open data. Most of the open data portals established by different governments are based on web-based platforms such as CKAN, Socrata, and OpenDataSoft (City of Vancouver, 2019a; Government of Canada, 2019; Socrata, n.d.). In the web environment more generally, users are likely to abandon systems with poor usability (Nielsen, 2012). Nielsen (2012) suggested that a useful system should have high utility and high usability. Utility refers to the effectiveness of features while usability is defined as a mixture of quality attributes including learnability, efficiency, memorability, error prevention, and satisfaction (Nielsen, 2012). A useful system sets the foundation to build up user engagement of technology (O’Brien & Toms, 2008). When users are engaged in technology, it may result in positive outcomes for e-learning, citizen inquiry, and participation (O’Brien, Cairns, & Hall, 2018). To support e-learning of data literacy skills through using an open data portal, the portal must have high utility and high. If the open data portal does not have functions to use data, such as filtering data and visualizing data, users may need to leave the portal and use external tools to make use of data. If an open data portal has low usability, users may be dissatisfied and abandon the portal.

There are some usability evaluation frameworks to help improve the usability of open data portals, and learning support is one of the common key aspects in evaluating the usefulness of open data portals (Máchová et al., 2018; Zhu & Freeman, 2019). Apart from searching for datasets, users would like to analyze, use, and discuss the datasets within the open data portal (Osagie et al., 2017). Osagie et al. (2017) found that users faced obstacles when finding, analyzing, and/or using datasets within the open data portal, and hence they would like to have
support from the portal to guide them on what to do. Help features are essential in designing open data portals.

Máchová et. al. (2018) conducted a broad review of best practices for improving the usability of open data portals and developed a list of criteria for evaluating usability organized along 3 dimensions: “open dataset specifications”, “open dataset feedback” and “open dataset request”. All the criteria within the dimension of “open dataset feedback”, including “documentation and tutorials”, “forum and contact form”, “user rating and comments” and “social media and sharing”, are related to learning activities such as reading, asking questions, and comparing (Máchová et al., 2018). Zhu and Freeman (2019) developed a user engagement evaluation framework of open data portals based on a review of the literature. There are several criteria within their user engagement evaluation framework related to learning such as user support, app showcase, and participation (Zhu & Freeman, 2019).

The design of open data portals can be guided by these usability evaluation frameworks, yet these theoretical frameworks which were developed through literature reviews (Máchová et al., 2018; Zhu & Freeman, 2019) may not result in user-centered designs that fit citizens with varying levels of data literacy skills. The design of open data portals has been criticized as overly developer-centric, and not designed for non-technical users (Osagie et al., 2017). It is important to understand how different types of users interact with open data and open data portals to generate usable open data portals across a wide range of users. The majority of literature evaluates open data portals by experienced data users and focuses on the principles of the open data portal, in which datasets and data portals should be open by default, timely and comprehensive, accessible and usable, comparable and interoperable, for improved governance and citizen engagement; and for inclusive development and innovation (Kubler, Robert,
Neumaier, Umbrich, & Le Traon, 2018; Maali, Cyganiak, & Peristeras, 2010; Umbrich, Neumaier, & Polleres, 2015; Vetrò et al., 2016). There is very limited research on understanding users’ interactions with open data portals, especially users who have little experience in working with data. In one study, researchers interviewed professionals who use data in their day-to-day work to learn how the professionals work on data-centric tasks. The study also analyzed participants’ search logs using the UK open data portal (Koesten, Kacprzak, Tennison, & Simperl, 2017). The study found that the majority of participants performed keyword searches with a single query and looked for datasets in structured data formats such as HTML, CSV, XLS, and JSON in the UK open data portal. The study did not go further to examine how participants interacted with the open data portal after searching for datasets (Koesten et al., 2017).

Integrating learning tools into an open data portal may support users to find and use open data within the portal when they encounter difficulties. In this study, learning tools are defined as any built-in features within an open data portal that can support users to learn to make use of open data. With a limited understanding of how users interact with open data portals, how different learning tools should be provided, and in what format within the open data portals to support learning remains unclear. Knowing the nature of informal learning provides a theoretical foundation for designing learning tools within open data portals.

2.4 Informal learning

Learning can happen informally when someone engages in tasks, in which learning is not primarily focused but a by-product through performing tasks. When someone is involved in formal learning such as learning in a classroom, taking an online course, or attending a training workshop, he is fully aware of himself as a learner in a formal learning process and he expects certain learning outcomes. In contrast, when someone performs a task, he informally learns or develops his skill
or knowledge for accomplishing the task without being fully aware of himself as an informal learner (Watkins, Marsick, Wofford, & Ellinger, 2018). In this situation, informal learning is invisible and embedded in the task performed (Vakkari, 2016). Informal learning is defined as learning that is not under the control of formal education institutions and lacks characteristics of formal learning, such as having an instructor, structured content, clear learning objectives, formal assessment, etc. (Hager & Halliday, 2009).

To acquire skills and knowledge through informal learning, learners need to go through a learning cycle which involves two main components: doing the task and reflecting on the experience (Boud, Keogh, & Walker, 1985; Davies, 2016; Jarvis, 1994; Kolb, 1984; Watkins et al., 2018). Among various informal learning models, Kolb’s experiential learning cycle is one of the best-known. Learners acquire skills or knowledge through four discrete phases: concrete experience; reflective observation; abstract conceptualization; and active experimentation (as shown in figure 2.1). Informal learners can begin at any stage and follow the consecutive stages to complete the whole informal learning cycle. For example, if someone wants to throw a ball far away, he/she can begin at the concrete experience stage to throw a ball, or he/she can begin at the reflective observation stage to watch someone else who throws a ball far away, or he/she can begin at the abstract conceptualization stage to read a book that teaches you how to throw a ball, or he/she can begin at the active experimentation stage to write up steps on throwing a ball.
For example, information searching can be an informal learning process to develop information searching skills. Vakkari’s information search process (2016) as a non-linear process involves three search stages and an information use stage: search formulation, selecting sources, interacting with sources, and synthesizing and presenting (as shown in figure 2.2). Different learning outcomes occur at different search stages, for instance, searchers learn new terms in the stage of search formulation; searchers learn the background of the topic in the stage of source selection; and searchers generate ideas in the stage of interacting with sources (Vakkari, 2016).
Although informal learning happens when individuals are involved in doing tasks, learning outcomes may not be positive. In an informal learning environment, learners may not have sufficient skills and knowledge to carry out tasks correctly and hence they may reach the wrong solutions and acquire misconceptions (Watkins et al., 2018). To help informal learners avoid acquiring inappropriate knowledge, supports and guidance should be provided within the informal learning environment so that learners can compare and reflect on their experience with a correct solution.

When individuals do tasks that involve planning, doing, reflecting, and concluding the experience, they learn something informally. Without enough support, individuals may not be able to perform tasks correctly and may lead to negative learning outcomes. Developing learning tools that can support stages of an informal learning cycle within a system may create a learning environment to help users to acquire skills and knowledge when using the system.

2.5 Research questions

In order to make use of open data within an open data portal, the open data portal should have high utility and usability while users should have sufficient data literacy skills. Some studies focus on improving the usability of open data portals (Máchová et al., 2018; Osagie et al., 2017; Weerakkody, Irani, Kapoor, Sivarajah, & Dwivedi, 2017) while some studies focus on developing users’ data literacy skills through formal training (Gascó-Hernández et al., 2018). If users do not have needed data literacy skills, users cannot make good use of open data even if the open data portal has high utility and usability. If the open data portal has poor utility and usability, it discourages users to make use of open data. To promote the use of open data, governments should improve the usability of open data portals and provide training to users to develop data literacy skills. Formal instructions may not be effective and efficient to reach out to a larger user group
due to time, location, and resource constraints. Building upon the notions of informal learning, would it be possible to develop learning tools within open data portals to support people to learn data literacy skills informally while they are using open data in open data portals? This study seeks answers to the following questions:

1. Do users with different levels of data literacy skills interact with an open data portal and associated tools and technologies to use open data to accomplish a task differently? If yes, how do they differ?

2. What obstacles due to data literacy do users encounter when using open data to accomplish a task?

3. What, if any, learning strategies do users employ when they meet obstacles in using open data to accomplish a task?
Chapter 3: Research Methods

3.1 Overview of research design

This study adopts the design thinking model and it focuses on the first ‘empathy’ stage of the design thinking process. The design thinking model is recommended by the Hasso-Plattner Institute of Design at Stanford for use in the design of user-centered products. The model has five stages, including empathize stage, define stage, ideate stage, prototype stage, and test stage. The main goal at the empathize stage is to understand what users do and how they do it through engaging users into a relevant context. The methods used to engage users may involve observation and interviews. The findings from the empathize stage provide key information to the next define stage in the design thinking model (Institute of Design at Stanford, n.d.).

This study is a remote qualitative study that involves a think-aloud observational session, a self-completion questionnaire, and a short semi-structured interview via Zoom videoconferencing. To find out how users with different levels of data literacy skills interact with an open data portal and associated tools and technologies, simulated work tasks were given to participants and the whole session was recorded for a play-back observation. When participants were working on the tasks, they were asked to think-aloud in order to capture their approaches to the tasks and record what obstacles they faced. The self-completion questionnaire collected relevant demographic information such as age range, educational background, and experience in using open data from the participants as well as measures of their self-perceived data literacy competence. The semi-structured interview captured participants' reflections on the obstacles they faced and the learning strategies they employed when working on the simulated work tasks, which may or may not be captured in the think-aloud session.
3.2 Selection of an open data portal

A system with good usability, usefulness, and user engagement may lead to positive outcomes for e-learning (O’Brien et al., 2018). Thus, it is essential to select an open data portal with high usability and interactivity. To ensure relevance to study participants who were residents of British Columbia, open data portals provided by the provincial government and municipal governments in British Columbia were considered. Potential open data portals were identified through a web search. Eight open data portals were examined, including DataBC provided by the Government of British Columbia and the open data portals of the following cities: Vancouver, Surrey, Nanaimo, Burnaby, Maple Ridge, White Rock, and Victoria. All of these have searchable catalogues to locate datasets and data explorer features to view datasets.

From among these candidates, the City of Vancouver’s (CoV) open data portal1 was selected for the study because it had the most comprehensive functionalities such as searchable catalogue, data tables with filtering and sorting functions, and data visualization tool. The CoV’s open data portal is the only open data portal that provides a customizable data visualization tool to analyze datasets within the open data portal, a component essential for this study. A new and enhanced version of the CoV’s open data portal was introduced shortly before this study began, and therefore the portal could also be considered state of the art.

We examined the selected CoV’s open data portal further to assess its usability and user interaction features based on criteria for open data portals (Máchová et al., 2018) and the user engagement framework for open data portals (Zhu & Freeman, 2019). The results of these analyses are presented in tables 3.1 and 3.2. The overall scores of the CoV’s open data portal

1 https://opendata.vancouver.ca/pages/home/
were higher than the mean score of assessed open data portals from Máchová et al. (2018)’s study and Zhu et al. (2019)’s study. Due to the comprehensive feature set, and high usability and high user engagement scores, the CoV’s open data portal was selected for this study.

Table 3.1 Evaluation of usability of CoV’s open data portal.

<table>
<thead>
<tr>
<th>Usability criteria</th>
<th>CoV’s open data portal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Máchová et al., 2018, p. 256)</td>
<td>Score (3 = done, 2 = partially fulfilled, 1 = unfulfilled) and explanation</td>
</tr>
<tr>
<td><strong>Criteria</strong></td>
<td><strong>Mean score (Canada’s open data portals)</strong></td>
</tr>
<tr>
<td>Description of dataset</td>
<td>3.00</td>
</tr>
<tr>
<td>Publisher of dataset</td>
<td>2.54</td>
</tr>
<tr>
<td>Thematic categories and tags</td>
<td>2.62</td>
</tr>
<tr>
<td>Release date and up to date</td>
<td>2.69</td>
</tr>
<tr>
<td>Machine-readable formats</td>
<td>2.54</td>
</tr>
<tr>
<td>Open data license</td>
<td>2.55</td>
</tr>
<tr>
<td>Visualization and analytics tools</td>
<td>1.92</td>
</tr>
<tr>
<td>Documentation and tutorials</td>
<td>2.62</td>
</tr>
<tr>
<td>Forum and contact form</td>
<td>1.62</td>
</tr>
<tr>
<td>User rating and comments</td>
<td>2.08</td>
</tr>
<tr>
<td>Social media and sharing</td>
<td>2.46</td>
</tr>
<tr>
<td>Request form</td>
<td>2.38</td>
</tr>
<tr>
<td>List of requests</td>
<td>2.08</td>
</tr>
<tr>
<td>Dimension</td>
<td>Measurement items</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>Can users browse datasets by categories                                                                ategori</td>
</tr>
<tr>
<td></td>
<td>Can users see a list of all available data sets?</td>
</tr>
<tr>
<td></td>
<td>Can users filter and/or sort these datasets?</td>
</tr>
<tr>
<td></td>
<td>Can users search for datasets?</td>
</tr>
<tr>
<td></td>
<td>Does the portal provide advanced search options (for instance, Boolean, field search, results filtering, and/or results sorting)?</td>
</tr>
<tr>
<td></td>
<td>Do users NOT need to sign up to use the data?</td>
</tr>
<tr>
<td></td>
<td>Is there an explicit cc-by or equivalent license?</td>
</tr>
<tr>
<td></td>
<td>Is the portal presented in the language(s) other than English?</td>
</tr>
<tr>
<td></td>
<td>Are all the data sets available in machine-processable format(s)?</td>
</tr>
<tr>
<td></td>
<td>Are all the data sets available in at least one open format?</td>
</tr>
<tr>
<td></td>
<td>Are all data sets available in multiple open formats?</td>
</tr>
<tr>
<td></td>
<td>Do none of the data sets include any .exe file?</td>
</tr>
<tr>
<td></td>
<td>Does each data set have a permanent URI?</td>
</tr>
<tr>
<td><strong>Total score</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Trust</strong></td>
<td>Is the total number of datasets available?</td>
</tr>
<tr>
<td></td>
<td>Does the portal include the most common categories—health, public safety, and transportation?</td>
</tr>
<tr>
<td></td>
<td>Does the portal differentiate between various types of data (such as datasets, documents, forms, calendars, and so on)?</td>
</tr>
<tr>
<td></td>
<td>Has anything been updated or added in the past 3 months?</td>
</tr>
<tr>
<td></td>
<td>Is the data policy available on the portal?</td>
</tr>
<tr>
<td></td>
<td>Can all the datasets be traced back to their primary source (such as specific government agencies who produced the data)?</td>
</tr>
<tr>
<td></td>
<td>Do all of the datasets include data more granular than city-level data?</td>
</tr>
<tr>
<td><strong>Is there any indication of how many times each dataset has been downloaded?</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>Is there any indication of how many times each dataset has been accessed?</strong></td>
<td>0</td>
</tr>
<tr>
<td><strong>Total score</strong></td>
<td><strong>17.78</strong></td>
</tr>
</tbody>
</table>

**Understand**

| **Is help information/FAQ available on the portal?** | 1 |
| **Is there contact information for users to contact portal managers and/or data set managers?** | 1 |
| **Are there any app showcases/examples available?** | 0 |
| **Are any of the data sets accompanied by supporting documentation specifying the data collection methods?** | 1 |
| **Are any of the data sets accompanied by a detailed data dictionary with definitions of terms/variables/fields?** | 1 |
| **Is metadata available for every data set?** | 1 |
| **Total score** | **16.67** |

**Engage**

| **Does the portal provide analytics about how data or the portal is used/accessed?** | 0 |
| **Is API present on the portal?** | 1 |
| **Does the portal suggest a citation format?** | 0 |
| **Can users personalize/customize the portal to allow features such as saving, search history, and so on?** | 1 |
| **Are all data sets downloadable?** | 1 |
| **Is a print option provided for each dataset?** | 0 |
| **Does the portal provide a batch download option?** | 0 |
| **Does the portal provide tools for online manipulation such as filtering and sorting within every dataset?** | 1 |
| **Does the portal provide map or other visualization tools for any of the datasets?** | 1 |
| **For any dataset, does the portal provide comparative datasets or recommend related data sets?** | 0 |
| **Total Score** | **10.00** |

**Participate**

| **Is there proactive engagement with the community (newsletters/e-mails/podcasts/social media)?** | 1 |
| **Are there tools available for the community to share datasets?** | 1 |
| **Can the community comment/discuss/rate data sets?** | 0 |
| **Does the portal ask users to suggest changes to existing data sets and/or suggest new data sets?** | 1 |
| **Total Score** | **15.00** |

**Overall**

| **Total Score** (Mean score from Zhu et. al.’s research finding: 64.81) | **76.37** |
3.3 Task and dataset selection

Participants were given simulated work tasks to find out how they interact with open data. Simulated work tasks control and produce conditions to all the participants for comparison, yet, they are artificial and hence participants may not have a context for executing the tasks or maybe unmotivated to work on the artificial tasks (Kelly, 2009). Citizens usually search for government information for life-events such as graduation, employment, and marriage (Haraldsen, Stray, Päivärinta, & Sein, 2004). Therefore, a naturalistic task in the open data context should be related to the life events of the participants. Since we planned to recruit university students who are or will be interested in their future careers, we considered that creating tasks related to careers and salary may motivate participants.

The Employee and Remuneration and Expenses Earning over $75,000 (ERE-75K) dataset\(^2\) was selected as a target dataset for the simulated work tasks. When searching for “salary” in the CoV’s open data portal, three datasets are retrieved. These include the ERE-75K dataset, the Council Budget and Expenses (CBE) dataset\(^3\), and the Workforce Pay Rates and Gender (WPRG) dataset\(^4\). The CBE dataset was not selected as a target for the study as the ‘Analyze’ tool, which is a customizable data visualization tool, is not available for use with it and the microdata cannot be explored within the portal. The WPRG dataset was not selected as participants need specialized knowledge to understand the data, such as ‘Admin Management I’ and ‘Admin Management Ia’ in the ‘Classification’ data variable, and ‘CUPE Local 15’ and ‘CUPE Local 1004’ in ‘Union’ data variable. The ERE-75K dataset was selected as the data is


\(^3\) [https://opendata.vancouver.ca/explore/dataset/council-budget-and-expenses/](https://opendata.vancouver.ca/explore/dataset/council-budget-and-expenses/)

relatively self-explanatory, such as ‘Firefighter’ and ‘Working Foreman’ in the ‘Title’ data variable, and the ‘Analyze’ tool is available in this dataset.

The simulated work tasks were designed to make use of the ERE-75K dataset in preparing a summary of findings to send to a friend. The work tasks aim to identify obstacles, if any, that participants may face due to limitations in data literacy. In order to allow participants to engage in the full cycle of informal learning (Kolb, 1984), we asked them to carry out two similar tasks, assuming they could apply what they learned in the first task when completing the second.

To fit within the 30-minute time limit of the think-aloud observational session, only two core competencies of data literacy, finding data and using data, were chosen to design the simulated work tasks. In order to find data, participants ‘need to be aware of the possible data sources, be able to evaluate them and select the ones most relevant to an informational need or a given problem, and be able to detect when a given problem or need cannot be totally or partially solved with the existing data’ (Calzada Prado & Marzal, 2013, p. 130). In order to use data, participants ‘need to be able to prepare data for analysis, analyze them in keeping with the results sought, know how to use the necessary tools, synthesize and represent the results of data analysis in a way that is suitable to the nature of data and the audience targeted in the inquiry, make ethical use of data, acknowledge the source, and make sure that results interpreted transparently and honestly’ (Calzada Prado & Marzal, 2013, p. 131). Simulated work tasks are comprised of a simulated work task situation and an indicative request (Kelly, 2009). It is recommended to “tailor the simulated work task situations towards the information environment and the group of test persons” (Borlund, 2003, p. 11), hence the simulated work task situation is designed so that university students are likely to relate to it. The simulated work task situation is designed as follows:
Imagine that you have a friend who is graduating from his/her undergraduate study soon. He/she is asking for your advice on his/her future career. You know that he/she may be interested in being a librarian or another occupation. You think it will help to let him/her know how much these two occupations earn. The City of Vancouver releases data on employee salary (also known as remuneration or earnings) in their open data portal, and you think it may be a good source to get information. You will reply to your friend with your findings in a table(s) or graph(s).”

The indicative request is designed to have participants write a reply about the salary of librarians and one other occupation in the last five years to the friend mentioned in the simulated work task situation. There are three sub-tasks of the indicative request as follows:

Task 1: Locate the needed dataset in the City of Vancouver’s open data portal

Task 2: Make use of located dataset to analyze the last 5 years salary (also known as remuneration or earnings) of

a. librarians and;

b. one occupation that your friend may be interested in (e.g. planner, analyst, etc.)

Task 3: Write your reply using the reply template

Task 1 is designed to identify obstacles that participants may face due to skill gaps in finding and selecting the most relevant dataset to solve the given problem. For example, participants may fail to select the ERE-75K dataset to answer the given problem. Task 2 and Task 3 are designed to identify obstacles that participants may face due to skill gaps in using data, includes wrangling data, analyzing data, reporting data, and citing data sources. For example, participants may fail to wrangle data when necessary; participants may fail to cite the used data sources in the reply. The study instructions given to participants are presented in Appendix A.1.
Participants may face obstacles in selecting the most appropriate dataset to answer the given problem if they lack data literacy skills. Therefore, the tasks should be designed in a way that only one dataset in the CoV’s open data portal can solve the given problem. There are over 160 datasets in CoV’s open data portal, and three of them are related to salary, including the CBE dataset, the ERE-75K dataset, and the WPRG dataset. The tasks are designed for the participants to make use of only the ERE-75K dataset. The WPRG dataset does not include information from Vancouver Public Library (VPL) and it only contains 2019 data. CBE dataset does not contain salary data other than mayor and councilors. Hence, participants are requested to analyze librarian salary for the last 5 years which can only be answered by making use of the ERE-75K dataset. Analyzing the salary of any selected occupation for the last 5 years creates a situation that participants may find interesting and may meet their own real information needs, which are recommended for designing simulated work tasks (Borlund, 2003). If participants select any dataset other than the ERE-75K dataset, it indicates that participants face challenges in selecting the most appropriate dataset.

3.4 Study procedures and instruments

Due to curtailment of on-campus research at the University of British Columbia during the COVID-19 pandemic (UBC Broadcast, 2020), a remote study was conducted in which the participants were required to use their computer equipment and to share their screen activities with the researcher via Zoom videoconferencing.

3.4.1 Think-aloud observational session

The 30-minute think-aloud observational session was designed to capture user interactions with open data and obstacles faced due to users’ knowledge gaps. To minimize fatigue to participants, the think-aloud observational session was limited to 30 minutes. Dunning
(2011) argued that people with low levels of expertise are often unaware of their inability and rate their skills highly. Hence, while data literacy was self-assessed by participants, the observation and the think-aloud session provided an opportunity to see them carry out data-related tasks and to observe different levels of skills more objectively. Details of instructions to participants and prompt used to encourage verbalizations are included in Appendix A.1.

3.4.2 Semi-structured interview

The semi-structured interview was designed to capture or provide more detail on the obstacles that participants faced and the learning strategies they took during the session. The interview lasted for no more than 15 minutes. The interview session consisted of 5 to 6 standard questions as follows (the script can be found in Appendix A.3):

1. First, how did you find the overall experience of using the portal?
2. Can you talk about any challenges or obstacles that you faced when completing the task?
   
   If the participant replied that he/she faced challenges or obstacles, the following four questions were asked:
3. What did you do to overcome the obstacles? Did you have any strategies to do that?
4. How helpful or unhelpful were those strategies?
5. Can you think of any features that could be built into the portal to help you overcome these obstacles?
6. At last, how do you think the open data portals can support people to learn data skills, especially those who may be beginners in working with data?

If the participant replied that he/she did not face any challenges or obstacles, the following three questions were asked:
3. Could you talk about any strategies you used that helped you to be successful in this task?

4. Can you think of any features that could be built into the portal that would have been helpful for you, in completing this task, or in using it in general?

5. At last, how do you think the open data portals can support people to learn data skills, especially those who may be beginners in working with data?

3.4.3 Study procedures

The study was carried out with participants individually via a password-protected Zoom videoconferencing session of up to 1 hour. The consent form of the study was sent to participants at least 24 hours before the study session. During the study session, the participant shared his/her computer screen with the researcher. Once the screen sharing was set up, the participant was asked to give verbal consent to participate in the study. After receiving the participant’s verbal consent, the participant first worked on a maximum 30-minute think-aloud observational session, then self-completed an online questionnaire, and finally answered a few questions in a semi-structured interview. The whole study session was recorded by Camtasia. The study instructions were distributed to the participant via a study website\(^5\). The details of the online questionnaire and the researcher’s script during the study session are shown in Appendix A.2, and Appendix A.3 respectively.

3.5 Recruitment of participants

University students are the selected population as they are active learners who may provide informative feedback on learning innovations. We also assumed that post-secondary students

\(^5\) [https://blogs.ubc.ca/ogdresearch/](https://blogs.ubc.ca/ogdresearch/)
would have sufficient levels of computer literacy to enable them to engage with the assigned
tasks. Fourteen participants were recruited using a convenience sampling approach between
April 30, 2020, and May 11, 2020. A recruitment notice was posted to the ‘Paid Participants
Studies Lists’ webpage⁶ which is hosted on the Psychology Graduate Student Council,
University of British Columbia (UBC) website. Participants meeting the following criteria were
included:

- UBC student and at least 17 years old;
- fluent in English; and
- access to a laptop or a desktop computer with a microphone, a web camera, Internet access,
  and a web browser installed.

3.5.1 Classifying participants based on their data literacy levels

Participants were classified into two groups, self-assessed higher data literacy group (SA-
HDLG) and self-assessed lower data literacy group (SA-LDLG), based on a self-assessment
survey. The data literacy assessment tool was created for this study, based on Calzada Prado and
Marzal’s framework on the core competencies of data literacy (2013). We identified statements
that typified each of the components in these frameworks and participants were asked to self-
assess their data literacy skills by indicating their level of agreement on a 5-point Likert scale
(i.e. 1-strongly disagree, 2-disagree, 3-neither agree nor disagree, 4-agree, and 5-strongly agree).
An online self-completion questionnaire was created in Qualtrics, as shown in Appendix A.2.
The instrument contains the following fifteen statements related to data literacy competency (as
shown in figure 3.1):

⁶ https://gsc.psych.ubc.ca/resources/paid-studies-list/
Figure 3.1 Data literacy assessment

For each statement, please use the following scale to indicate your level of agreement:

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I can identify different kinds/types of data</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can describe how data is produced in society</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can describe a wide range of possible applications of data</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can find credible data to solve a given problem or need.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can evaluate the quality of data</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can identify the limitations of data for a given problem or need.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can describe the various ways to represent data (e.g. written, numerical or graphic)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can interpret data in the form of charts or tables</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can describe data with metadata</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can manage data for subsequent reuse</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can clean and prepare data for analysis</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can analyze data using tools such as spreadsheets or statistical software</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can summarize the results of data analysis using various forms of representation (e.g. written, numerical or graphic)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can use data ethically (e.g. protect confidentiality, data license)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>I can cite data sources I have used.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
For those participants who self-assess at least one statement as 1-strongly disagree or 2-disagree are grouped as SA-LDLG while for those participants who self-assess all the statements as 3-neither agree nor disagree, 4-agree, or 5-strongly agree are grouped as SA-HDLG. An online self-completion questionnaire was created using the Qualtrics Survey Tool to implement the Data Literacy Assessment, as shown in Appendix A.2.

3.6 Data Analysis

3.6.1 Processing Camtasia recordings

The Camtasia recordings of the think-aloud observational sessions and semi-structured interviews were imported into NVivo 12 for transcription. For the think-aloud observational sessions, the recordings were split into multiple entries (i.e. Entry) in NVivo when the participant visited a page, used an application, or worked on a task. The Timespan of each Entry is captured automatically in NVivo. Apart from Timespan, each Entry contains Transcript and Actions. What the participant said was transcribed into the Transcript column and what the participant did was described into the Actions column of each Entry in NVivo. An extract of transcript entries of a think-aloud session of a participant is shown in table 3.3.
Participant N01’s extract of transcript entries of the think-aloud session

<table>
<thead>
<tr>
<th>Entry</th>
<th>Timespan</th>
<th>Transcript</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>00:00.0 - 00:32.9</td>
<td>I am going to use the open data portal now. I am trying to figure out the amount… the salary of a librarian for my friend to use. Yeah, I am trying to see where I can find the data. As this is the first time I use the site, so I am trying to get a holistic view of the site.</td>
<td>Browsed the homepage of the CoV open data portal</td>
</tr>
<tr>
<td>2</td>
<td>00:32.9 - 00:38.1</td>
<td>(silent)</td>
<td>Typed &quot;library&quot; onto the search bar on the homepage of the portal</td>
</tr>
<tr>
<td>3</td>
<td>00:38.1 - 00:48.0</td>
<td>Ok. So, the first one is the workforce pay rates and gender. And I think this is the one that is relevant in this case.</td>
<td>3 results were retrieved, including the WPRG dataset, “awarded contracts” dataset, and &quot;libraries&quot; dataset. Browsed the results and then selected the WPRG dataset.</td>
</tr>
</tbody>
</table>

For the semi-structured interviews, Camtasia recordings were split into multiple transcript entries (i.e. Entry) in NVivo when the participant answered a new question. An extract of transcript entries of a semi-structured interview of a participant is shown in table 3.4.

<table>
<thead>
<tr>
<th>Entry</th>
<th>Timespan</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>00:00.0 - 00:22.7</td>
<td>[Researcher: First, how did you find the overall experience of using the portal] It was ok once I knew what I was looking for. I kind of struggled in the beginning. Because I typed salary of librarian or librarian-ish term, like very specific terms that won't come up from the general portal. When I typed (general term) like earnings, then everything is much clear.</td>
</tr>
</tbody>
</table>

Camtasia recordings were selectively transcribed, focusing on instances when he/she faced obstacles, such as saying “I think I am doing incorrectly” and doing something incorrectly, and when he/she tried to overcome the obstacles, such as saying “I will try this one instead” and looking for help.
3.6.2 Analyzing entries

The entries of think-aloud observational sessions and semi-structured interviews were coded in NVivo based on four main themes: activities, tools used, obstacles, and learning strategies. The codes for the first three themes: activities, tools used, and obstacles, were pre-defined based on the research questions, while the codes for the ‘learning strategies’ theme were developed during data analysis. The following paragraphs indicate how codes were derived.

The activities were coded using two pre-defined codes: (a) find data, and (b) use data according to the definitions identified in the literature (Calzada Prado & Marzal, 2013). To succeed in task 1, which is related to find data, participants should perform dataset searching in the CoV’s open data portal, select the ERE-75K dataset among all the retrieved datasets, detect that other datasets such as the WPRG dataset and the CBE dataset cannot solve the given problem, and detect the limitations of the ERE-75K dataset, namely that it only captures employees who earned over $75,000 per year in which some occupations with starting salary lower than $75,000 were not included. To succeed in task 2 and task 3, which are related to using data, participants must extract remuneration data of selected occupations from 2015 to 2019 from the ERE-75K dataset, wrangle data when necessary, analyze the cleaned data using the CoV’s open data portal or other data analysis tools such as Microsoft Excel, create an accurate table or graph, write down how the results were analyzed, and cite the used data source.

The tools and webpages used were coded using two sets of pre-defined codebooks according to the available components in the CoV’s open data portal. Codebook 1 is designed for coding the “find” data activities while Codebook 2 is designed for coding the “use” data activities. Codebook 1 focuses on the major tools and components within the CoV’s open data portal such as the ‘Homepage’ and the ‘Help & API’ page. Codebook 2 focuses on the major
tools within the dataset page in the CoV’s open data portal such as the ‘Table’ tool, which allows users to filter and sort the dataset, and the ‘Analyze’ tool, which allows users to visualize the dataset. All the codes for *Codebook 1* and *Codebook 2* are shown in Appendix B.1 and Appendix B.2 respectively.

Obstacles were classified into two categories: (a) recognized obstacles and (b) unrecognized obstacles. If a participant noticed that he/she was facing an obstacle, the obstacles were classified as recognized obstacles. For example, a participant expressed that he/she did not know how to use filter data in Microsoft Excel, the expressed obstacle was classified as a recognized obstacle. If a participant did not notice that he/she was facing an obstacle, the observed obstacles were classified as unrecognized obstacles. For example, a participant selected a non-ERE-75K dataset and did not notice the dataset cannot answer the given problem, the observed obstacle was classified as an unrecognized obstacle. The obstacles were also coded using a top-down approach, in which the top-level codes were developed based on Calzada Prado & Marzal’s (2013) data literacy competency framework. The pre-defined codes are listed in *Codebook 3* in Appendix B.3.

The learning strategies were coded using a bottom-up approach, in which the codes were developed when analyzing the entries. When a participant faced recognized obstacles, he/she might take different learning strategies to try to overcome the obstacles. The action taken to overcome obstacles was extracted. Then all the extracted actions were grouped based on their similar nature to develop codes as participants’ learning strategies. For example, a participant tried keyword “librarian” and “salary” to search for datasets while another participant tried keyword “librarian salary” and ‘salaries’ to search for datasets, both of their actions involved trying different approaches until some relevant datasets were retrieved, their actions were coded
as trial-and-error; When a participant did not know how to apply a filter in Excel and then he/she searched for instructions via Google search engines, the action taken was coded as looking for instructions.

3.6.3 Processing self-completion survey

The data of the self-completion questionnaire was extracted from Qualtrics and imported to Microsoft Excel for data processing. Each participant was classified into either SA-HDLG (high literacy) or SA-LDLG (low literacy), according to their self-assessment of data literacy score. The demographics of each participant are merged into entries for data analysis. The summary of demographics among two self-assessed data literacy groups is listed in table 3.5. The demographic profile of each participant is shown in Appendix C.1.
Table 3.5 Demographics of participants compared with their self-assessed data literacy groups

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Total</th>
<th>Self-assessed data literacy group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SA-HDLG</td>
</tr>
<tr>
<td>Study level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undergraduate</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Master</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Subject discipline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arts (incl. Faculty of Arts, and Faculty of Education)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Sciences (incl. Faculty of Sciences, Faculty of Applied Science, and Faculty of Land and Food Systems)</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Age range</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17-21</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>22-26</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>27-31</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Self-assessed level of knowledgeable on open government data (OGD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No knowledge at all</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Very little knowledgeable about it</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Little knowledgeable about it</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Some knowledgeable about it</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Very knowledgeable about it</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Experience in using any OGD website</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>No</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Experience in using CoV’s open data portal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
<td>6</td>
</tr>
</tbody>
</table>

Of the 14 participants, 6 were grouped into SA-HDLG and 8 participants were grouped into SA-LDLG in this study. Ten participants were undergraduate students, and four participants were Master’s students. Eleven participants were from Sciences disciplines while 3 participants were from Arts disciplines. All participants were aged between 17 and 31. Over half of the participants rated themselves as having no or very little knowledge about OGD and only 3 participants had experience in using any OGD portals. None of the participants used the CoV’s
open data portal before the study. The self-assessment of the data literacy survey among each participant is shown in Appendix C.2.
Chapter 4: Research findings

This chapter contains three sections, and each section addresses one research question listed in section 1.4. Section 4.1 addresses the first research question and presents how participants from SA-HDLG and SA-LDLG interact with the CoV’s open data portal, open data, and other associated tools to accomplish the assigned tasks. Section 4.2 addresses the second research question and presents obstacles faced by participants, based on the Calzada Prado & Marzal’s data literacy competency framework. Section 4.3 addresses the third research question and presents learning strategies employed by participants when they faced obstacles.

4.1 Participants’ interactions

4.1.1 Participants’ interactions to find data

The majority of participants found datasets by keyword searching, and the minority of participants browsed for datasets in the CoV’s open data portal. A search bar is located at the center of the homepage of the CoV’s open data portal which allows users to type in any keywords to search for datasets. Below the search bar, users can browse datasets by themes, such as ‘Business and economy’ and ‘Government and finance’. The VanDashboard page in the CoV’s open data portal provides some performance indicators, such as ‘Rental vacancy rate’ and ‘Crime Severity Index’, which allow users to access the corresponding datasets. ‘Chart builder’ and ‘Map builder’ are built-in data visualization tools that allow users to visualize any datasets in the CoV’ open data portal into a graph or a map (City of Vancouver, 2019b). Keyword searching was used more for finding datasets than browsing. A summary of participants’ interactions within the CoV’s open data portal to find the ERE-75K dataset is shown in table 4.1. Five participants found datasets by browsing datasets by theme on the homepage of CoV’s open data portal, browsing indicators on the VanDashboard webpage of CoV’s open data portal, and/or
browsing datasets on the chart/map builder. However, they could not find the relevant datasets through browsing and so they performed keyword searching. Another nine participants found the dataset by keyword searching directly.

Table 4.1 Participants’ interactions with CoV’s open data portal to find ERE-75K dataset

<table>
<thead>
<tr>
<th>Tools used to find ERE-75K dataset</th>
<th>Number of participants (Participant number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoV’s open data portal – Main Page</td>
<td>Total</td>
</tr>
<tr>
<td>Homepage</td>
<td>14</td>
</tr>
<tr>
<td>Keyword search</td>
<td>14</td>
</tr>
<tr>
<td>Browsing datasets by theme</td>
<td>2</td>
</tr>
<tr>
<td>VanDashboard</td>
<td>3</td>
</tr>
<tr>
<td>Chart/Map builder</td>
<td>2</td>
</tr>
</tbody>
</table>

All but one participants (N06) were able to find the ERE-75K dataset. A summary of the webpages visited, and time spent to find the ERE-75K dataset among participants is shown in figure 4.1. Ten participants were able to search for salary-related datasets and select the ERE-75K dataset within 6 minutes to find the ERE-75K. These participants primarily used the homepage/catalogue of the CoV’s open data portal. Four participants (N05, N06, N09, and N12) spent over 16 minutes in navigating different pages in the CoV’s open data portal such as VanDashboard, chart/map builder, and irrelevant datasets. They also visited other websites outside the CoV’s open data portal to figure out where to look for datasets (as shown in figure 4.1). The datasets and other websites visited by participants are listed in Appendix C.3 and Appendix C.4 respectively.
4.1.2 Participants’ interactions to use data

The CoV’s open data portal and Microsoft Excel were used to analyze data. A summary of data analysis tools used to analyze the ERE-75K dataset is shown in figure 4.2. Eight participants used two to three different data analysis tools. Four participants primarily used the table tool in the CoV’s open data portal, six participants primarily used the ‘Analyze’ tool in the CoV’s open data portal, and three participants primarily used Microsoft Excel to analyze the ERE-75K dataset. The use of external tools (i.e. Excel) was more predominant among participants in the SA-LDLG, while those in the SA-HDLG primarily used the data within the CoV’s open data portal. Since participant N06 failed to locate the ERE-75K dataset, he was excluded when analyzing participants’ interactions with the ERE-75K dataset.
Many participants in the SA-HDLG were able to utilize the ‘Table’ tool and ‘Analyze’ tool in the CoV’s open data portal, hence none of them exported the dataset to analyze in another data analysis tool. Since the CoV’s open data portal can only calculate the average remuneration per year, two SA-HDLG participants (N01 and N13) used Microsoft Excel as a supplementary data analysis tool to calculate average remuneration over the last 5 years.

For the participants in the SA-LDLG, five participants (63%) exported the dataset from CoV’s open data portal to Microsoft Excel. Among these five participants, two of them found it difficult or ‘annoying’ (N02) to analyze the dataset in Microsoft Excel and hence they shifted back to use the ‘Table’ tool or ‘Analyze’ tool in the CoV’s open data portal to analyze the dataset. Three participants chose to use Microsoft Excel to analyze the ERE-75K dataset as they could not figure out how to use the ‘Table’ tool or ‘Analyze’ tool in the CoV’s open data portal. One participant (N10) in the SA-LDLG mentioned in the interview session that he chose to use Microsoft Excel because he did not know how to use the ‘Analyze’ tool in the CoV’s open data portal.

“The analytic tool [i.e. ‘Analyze’ tool in the CoV’s open data portal], I was completely lost. I did not know how to use that. I am not sure if there is any help feature that can guide
me to use the analysis tool. Maybe they didn’t have it. I would suggest they add one.”

(N10, SA-LDLG).

The other two participants (N05 and N12) in the SA-LDLG skipped the ‘Table’ tool or ‘Analyze’ tool in the CoV’s open data portal and exported the dataset in CSV/XLS format for data analysis in Microsoft Excel. These two participants spent over 22 minutes to locate the ERE-75K dataset and hence they were running out of time to analyze the ERE-75K dataset. So, they may not have had enough time to explore data analysis tools in the CoV’s open data portal which they never used before and chose to use Microsoft Excel with which they were familiar.

Participants were asked to write up the results of their analysis. Thirteen participants presented their findings in Microsoft Word and one participant wrote the findings in Apple Pages. One participant (N05) in the SA-LDLG and one participant (N06) in the SA-HDLG wrote the findings based on the WPRG dataset while the other participants wrote the findings based on the ERE-75K dataset. Although the participant (N05) in the SA-LDLG was able to select and analyze the ERE-75K dataset after 23 minutes of study session (as shown in figure 4.5), she ran out of time to write up the results of the analyzed ERE75K dataset. The participant (N06) in the SA-HDLG did not analyze the ERE-75K dataset. Eleven participants manually input analyzed data into their write ups by reading the data values on the graphs or tables in CoV’s open data portal, or the exported Excel files. Four participants used a snipping tool from their local computer to take a screenshot of the analyzed graph or table from the CoV’s open data portal to the word processing software. Two participants (N11 and N13) in the SA-HDLG cited the used data source by inserting a hyperlink to the ERE-75K dataset. One participant (N02) in the SA-LDLG cited the used open data portal by inserting the hyperlink to the CoV’s open data portal, but without referring to the ERE-75K dataset. The remaining participants did not cite the ERE-
75K dataset or the CoV’s open data portal. In general, the participants in the SA-HDLG performed slightly better based on the write ups than the participants in the SA-LDLG. Among the eight participants who finished the write ups, only two participants (N11 and N13), both in the SA-HDLG, properly labelled all the graphs/tables, wrote down the selected job titles being analyzed, and cited the data sources. Other participants in both SA-HDLG and SA-LDLG failed to label the graphs/tables, did not mention which job titles were being analyzed, or/and did not cite the data source.

4.1.3 Summary of participants’ interactions

To answer the first research question of “how do users with different levels of data literacy skills interact with an open data portal and associated tools and technologies to use open data to accomplish a task?”, the findings indicate that many participants in both SA-HDLG and SA-LDLG primarily used the CoV’s open data portal to find and analyze the dataset, while some participants in SA-LDLG used it to find the dataset only. Many participants in both SA-HDLG and SA-LDLG searched for datasets on the homepage of the CoV’s open data portal while some participants in both groups tried to locate the dataset through browsing the CoV’s open data portal, which was an unsuccessful strategy for this task. The use of external tools was more predominant among participants in the SA-LDLG, while those in the SA-HDLG primarily made use of the data within the CoV’s open data portal. There were no apparent differences between the two groups of participants in writing up the findings. Many participants manually copied the findings from the CoV’s open data portal into word processing software, mainly Microsoft Word. Some participants used the snipping tool as a supplementary tool to copy graphs from the CoV’s open data portal to Microsoft Word.
4.2 Obstacles faced

Participants faced different types of obstacles associated with data literacy and system usability. The study found that participants faced obstacles when finding data: searching for datasets and selecting the most appropriate dataset; and using data: analyzing quantitative data, and presenting analyzed findings. These are presented based on their activities in the sections to follow.

4.2.1 Finding a dataset

Some participants faced obstacles when finding datasets due to usability issues of the CoV’s open data portal and limited data literacy skills. The obstacles faced in sub-section 4.2.1.1 arose mainly due to usability issues while the obstacles faced in sub-section 4.2.1.2 and 4.2.1.3 seemed to be the result of participants’ limited data literacy skills.

4.2.1.1 Where to search

While the majority of participants found the CoV’s open data portal design was intuitive to use when searching for datasets as they located the search bar and then performed keyword searching on the homepage within a few minutes, some participants faced obstacles in navigating the CoV’s open data portal to search for datasets. Determining the relevant page and tools to search for datasets was a challenging task for one participant (N06) in the SA-HDLG and three participants (N05, N09, and N12) in the SA-LDLG. Although participant N06 self-assessed as having high data literacy skills, he was the only participant who faced extreme challenges in searching for datasets and failed to locate the ERE-75K dataset. As shown in figure 4.5, he spent a great deal of time browsing through different pages of the CoV’s open data portal. He expressed his difficulties in understanding what pages would be relevant:
“I am looking for some salary charts from this website, I don't know where I can find it...Actually, I don't know what this (i.e. VanDashboard) is right now...I don't know where to look for it (i.e. on the catalogue page).... Help and API, I don't know what it is.....I am not sure if this (i.e. map builder) is helpful, but I am looking for some job data....I think this website doesn't include any information regarding jobs or other related areas, it is all about business improvement...I don't know how to use this tool....what can I do?” (N06, SA-HDLG)

He spent approximately 13 minutes until he realized that he was unable to figure out how to use the CoV’s open data portal by himself and hence he asked the researcher for help. The researcher guided him in searching for the dataset so that he could move to the next task. Participant N05 in the SA-LDLG also faced difficulty in navigating the relevant webpages to search for datasets. She visited an external website, the VPL’s open data research guide, to look for a dataset.

“It is more difficult than I expected. I thought it will be straightforward. The salary should be somewhere (i.e. VPL’s open data research guide)” (N05, SA-LDLG).

Some participants found it hard to figure out where to search for datasets, which may reflect a lack of familiarity or a poor understanding of how data is shared through an open data portal – in the form of datasets, or what kinds of tools can be used to locate these datasets. This may have been made more challenging due to a large volume of information on the CoV’s open data portal, as every extra unit of information diminishes the relative visibility of the search bar on the homepage of the CoV’s open data portal. The usability issues of the CoV’s open data portal set obstacles to users in navigating the portal.
4.2.1.2 How searching works

The search system of the CoV’s open data portal matches the search query with the metadata of the datasets, which includes the title and search terms assigned to datasets, rather than to the full text of the microdata, which includes the headers of the data tables and data values. To retrieve relevant datasets, users need to formulate general search queries that reflect the nature of datasets rather than the specific data values they are seeking. For example, the metadata of the ERE-75K dataset includes ‘salary’ and ‘salaries’ as search terms but and it does not include “librarian’, while the microdata of the ERE-75K dataset contains ‘librarian’ as a data value in the ‘Title’ data variable and it does not include “salary” in the headers of the data table in the ‘Table’ tool. Since the catalogue of the CoV’s open data portal matches the search query with metadata rather than the full text of microdata, the search query “salary” can retrieve the ERE-75K dataset while the search query “librarian” cannot. This kind of metadata searching was unfamiliar to some participants in both SA-HDLG and SA-LDLG and hence they faced obstacles when selecting appropriate search queries to yield search results.

Figure 4.3 Number of search queries among participants
On average, a participant formulated 2 search queries. Two out of six participants (33%) in the SA-HDLG formulated 3 search queries while three out of eight participants (38%) in the SA-LDLG formulated 3 to 5 search queries. On average, fewer search queries were formulated by participants in SA-HDLG (2) than by those in the SA-LDLG (2.65). The number of formulated search queries is shown in figure 4.3.

Table 4.2 Summary of search queries among the order of attempts

<table>
<thead>
<tr>
<th>Search queries</th>
<th>Number of participants</th>
<th>Order of attempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total 1st 2nd 3rd 4th 5th</td>
<td></td>
</tr>
<tr>
<td>Employee</td>
<td>2 2</td>
<td></td>
</tr>
<tr>
<td>Employee earnings</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>Employee salary</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>Job</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>Last 5 years job</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>Librarian</td>
<td>3 2 1</td>
<td></td>
</tr>
<tr>
<td>Librarian salary</td>
<td>2 2</td>
<td></td>
</tr>
<tr>
<td>Librarians</td>
<td>3 1 2</td>
<td></td>
</tr>
<tr>
<td>Library</td>
<td>4 2 1 1</td>
<td></td>
</tr>
<tr>
<td>Pay rate</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>Remuneration</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>Remuneration or earnings</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>Salaries</td>
<td>2 1</td>
<td></td>
</tr>
<tr>
<td>Salary</td>
<td>8 5 1 1</td>
<td></td>
</tr>
<tr>
<td>Salary for librarian</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>Salary Vancouver</td>
<td>1 1</td>
<td></td>
</tr>
</tbody>
</table>

Half of the participants started by searching for a specific search term such as “library”, “librarian(s)”, or “librarian salary” as shown in table 4.2. Many participants realized that they faced an obstacle as no relevant datasets were retrieved. However, they did not know how searching for datasets in the CoV’s open data portal is different from searching for full-text documents in popular search engines like Google, which index all terms in a document. Two out of six participants (33%) in the SA-HDLG faced this obstacle while six out of eight participants (75%) in the SA-LDLG faced this obstacle. Some of them changed the search query in a
different form, for example from “librarian” in the first attempt to “librarians” in the second attempt. To overcome the obstacles, many of them found a clue from the task description on the study website and tried other potential search queries that appeared there, such as “salary”, “remuneration”, “employee”, or “remuneration or earnings”. Participant N01 reflected on this obstacle in the semi-structured interview, and how she changed the approach to search to yield search results:

“I kind of struggled in the beginning. Because I typed salary of librarian or librarian-ish term, like very specific terms that won't come up from the portal. When I typed general terms like earnings, then everything is much clearer.” (N01, SA-HDLG).

4.2.1.3 Selecting the most appropriate dataset

To select a dataset to solve a given problem, the retrieved dataset should be evaluated on its relevance, usability, and quality which may rely upon features, such as coverage, time frame, documentation, license, dataset schema, data format, and collection methods (Koesten et al., 2017). Data evaluation helps to detect whether a dataset can solve totally or partially the given problem and enables the selection of the most appropriate dataset among different relevant datasets. Some participants demonstrated limited data evaluation skills, which caused them to work on irrelevant datasets for a long time.

In the description of the WPRG dataset, it states the following:

“This dataset provides the ranges of hourly rates of pay for all City job classifications and the gender breakdown of staff within these classifications. The dataset does not include information from the Vancouver Public Library and the Vancouver Police Department” (City of Vancouver, 2019b)
Although the WPRG dataset description clearly states that it does not have any information from the Vancouver Public Library, nine out of fourteen participants (64%) selected the WPRG dataset which could not solve the given problem in this study. Five participants (36%) also selected other irrelevant datasets such as the CBE dataset and the Libraries dataset. All the datasets visited by participants are listed in Appendix C.3. Among nine participants who explored WPRG dataset, five of them (N01, N07, N08, N10, and N11) were able to determine that it was not suitable for this study within 2 minutes (as shown in figure 4.5) by evaluating the coverage and time frame of the WPRG dataset.

“I am expanding the dataset schema. I realize this [i.e. WPRG dataset] is about the pay rate of the general workforce between women and men, but not for jobs specifically. So, I am going to look at employee remuneration and expenses dataset.” (N10, SA-LDLG).

Three participants (N05, N09, and N12) in the SA-LDLG and one participant in the SA-HDLG (N06) analyzed the WPRG dataset for 7 to 30 minutes (as shown in figure 4.1). They determined that the WPRG dataset was useful by evaluating the relevance of the dataset based on the title of the dataset rather than evaluating the relevance, usability, and quality of the dataset based on the microdata and data documentation such as the data schema and data collection methods. They could not find any data related to librarians or data between 2015 and 2018 in the WPRG dataset. However, they still spent a long time looking for the needed data in WPRG dataset because they believed it was relevant. In at least two cases, this was due to a belief that the data was there, but they just did not know how to use the available tools to extract the needed data. They did not recognize that the obstacles they were facing were due to the knowledge gap of data evaluation rather than the knowledge gap of using the tools.
“I feel like it [is] actually in the Excel [i.e. CSV file of WPRG dataset], I just don’t know how to interpret it.” (N05, SA-LDLG).

“So I have 2019 data, and I want to add 2018 data [i.e. the WPRG dataset]. Maybe add a series….I will click on the admin management again to see how to get the last 5 years” (N09, SA-LDLG).

After asking the researcher for assistance, three participants (N05, N09, and N12) in the SA-LDLG decided to search for another dataset but one participant (N06) in the SA-HDLG continued to use the WPRG dataset until the study session ended. This obstacle was more prevalent among the lower data literacy group. Only one participant (17%) in the SA-HDLG did not evaluate whether the dataset can solve the given problem. Three participants (38%) in the SA-LDLG did not perform data evaluation when selecting the most appropriate dataset.

4.2.2 Using the dataset

Participants were required to analyze the selected dataset and then present the analyzed findings. Analyzing the dataset involves two main steps, data wrangling, and data analysis. Data wrangling, a process of iterative data exploration and transformation such as diagnosing data problems, correcting errors, and transforming unstructured data into structured data, is an important step before analyzing data (Kandel et al., 2011). Data analysis involves choosing appropriate data analysis methods and data analysis tools. Users need to have both statistical knowledge and knowledge of data analysis tools to analyze data properly. Statistical knowledge is the foundation to determine the appropriate data analysis method to analyze quantitative data. Depending on the types of data variables, different kinds of tables or diagrams should be used in quantitative data analysis (Bryman, 2012). Knowledge of data analysis tools is a prerequisite to carry out the chosen data analysis method. Many participants faced unrecognized obstacles due
to limited data literacy skills while some participants faced recognized obstacles when using built-in data analysis tools within the CoV’s open data portal due to usability issues.

4.2.2.1 Wrangling data

The ERE-75K dataset is in the form of a structured data table, which requires minimal data wrangling. However, the dataset does contain some inconsistent data due to multiple representations of data across different years which requires data cleaning. However, none of the participants performed data wrangling such as diagnosing data problems and cleaning the data.

Only one participant (N11) in the SA-HDLG noticed the multiple representations of data for “Planner Ii” and “Planner II” which refers to the same occupation across different years. However, she failed to merge these two multiple representations of data into a single consistent data variable for analysis. Instead, she created two separate graphs for “planner Ii” and “planner II” (as shown in figure 4.4 and 4.5 respectively).

Figure 4.4 Participant N11’s graph for ‘Planner Ii’ in the write up

![Planner Ii Graph](image-url)
Firefighters were under the department named ‘VFPRS & OEM’ since 2017 and the previous department name was ‘Fire and rescue services’ in the ERE-75K dataset. Participant N10 in the SA-LDLG did not realize the need for diagnosing data problems such as the multiple representations of the department in the ERE-75K dataset, even though such data issues are very common in open government data. Therefore, when he analyzed the salary for firefighter, he could not locate salary data for firefighters between 2017 and 2019 (as shown in figure 4.6) as he filtered data for the “Fire and rescue services” department only.

<table>
<thead>
<tr>
<th>Firefighter Salaries</th>
<th>Salary ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>75148.46</td>
</tr>
<tr>
<td>2015</td>
<td>75221.61</td>
</tr>
<tr>
<td>2014</td>
<td>75025.45</td>
</tr>
<tr>
<td>2013</td>
<td>75150.63</td>
</tr>
<tr>
<td>2012</td>
<td>75002.07</td>
</tr>
</tbody>
</table>

- Salaries are lowest of all firefighters in Fire and Rescue Services of the year
Participants from SA-HDLG and SA-LDLG demonstrated a similar lack of data literacy skills in data wrangling, as none of them tackled this issue. This may be due to a lack knowledge of the prevalence of errors and inconsistencies in open data, a lack data wrangling skills, or they may have lacked time to spend on data wrangling due to time constraints.

4.2.2.2 Analyzing data

The ‘Table’ tool and ‘Analyze’ tool in the CoV’s open data portal, and Microsoft Excel were the data analysis tools used by participants. All participants in the SA-HDLG were able to figure out how to use the ‘Table’ tool and/or ‘Analyze’ tool in the CoV’s open data portal. This success seems to be based on the participants in the SA-HDLG knowing how the data should be analyzed, so they simply needed to try different functions in the ‘Table’ tool and/or ‘Analyze’ tool in the CoV’s open data portal to find and use the tools to perform data analysis. Due to limited time, some participants in the SA-LDLG chose to use Microsoft Excel, with which they were more familiar, to analyze the dataset, rather than spending time learning how to use the unfamiliar tools in the CoV’s open data portal.

“There is a bit of the learning curve trying to figure out if I could filter the data, and if I could filter by department. I struggled to try to even figure out how to change the y- or x-axis.” (N01, SA-HDLG)

The participants in both the SA-HDLG and SA-LDLG faced challenges in using the ‘Analyze’ tool due to usability issues. A serious constraint was the lack of help features or documentation on the ‘Analyze’ tool to guide users to use different functions. As an example, participant N11 in the SA-HDLG wanted to create a line chart of average remuneration across years by selected job titles. She was able to figure out how to use filters, set variables for the x-axis and y-axis, set break down series, and choose the graph type as a line chart to create an
appropriate graph. She noticed that the line chart was stacked, but she did not how to unstack the line chart. Hence, she read the data value from the graph and then manually typed the data value into a table she created in the Microsoft Word document, rather than copying the graph directly (as shown in figure 4.17).

Figure 4.7 Participant N11’s stacked line chart and the write up

When she analyzed the second occupation, she filtered just one occupation to avoid creating a stacked line chart (as shown in figure 4.8).

Figure 4.8 Participant N11’s normal line chart
Another example, participant N02 in the SA-LDLG wanted to create a column chart to compare the average remuneration of two selected occupations. However, he did not know how to use the filter to select the needed titles which caused an error message. He read the error message but he could not figure out how to fix the error by applying filters that he wanted (as shown in figure 4.9). Hence, he analyzed average remuneration per year by department (as shown in figure 4.10).

“I am trying to find the average remuneration of each year for librarians...I can't get...

Ok, I guess...So now I am looking for Vancouver Public Library Board instead of librarians specifically...I found it a bit difficult to find the exact salaries for the specific librarian positions” (N02, SA-LDLG).

Figure 4.9 Participant N02’s graph by job titles
Some participants in the SA-LDLG faced obstacles in analyzing the data due to the lack of data literacy skills. They did not know how to process and analyze quantitative data. For example, participants N07 and N09 in the SA-LDLG did not know how data should be arranged in a graph in order to satisfy the task. Hence, they created meaningless graphs and analyzed the salary of selected occupations incorrectly (as shown in figure 4.11 and 4.12).
Figure 4.11 Participant N07’s graph
Due to the time limit, not all participants were able to complete Task 3 to present the analyzed findings for a friend based on the scenario. Those in the higher data literacy skills group were more successful: five out of six participants (83%) in the SA-HDLG and three out of eight participants (38%) in the SA-LDLG completed this part of the study. Some participants in the SA-HDLG wrote up their findings in detail so that the analyzed findings could be re-used by the friend based on the scenario. As an example, participant N13 indicated the data referred to the selected employee, created a table with proper titles and headers, and also provided a link to the data source (as shown in figure 4.13).
In most cases, the findings presented by participants in the SA-LDLG were less clear than those presented by participants in the SA-HDLG. As an example, participant N08 presented the salary range of ‘Librarian’ without indicating which titles were included in his analyzed findings. He only analyzed ‘Librarian I’, ‘Librarian II’, ‘Librarian III’, and ‘Librarian IV’ data and excluded the data for all other librarians such as ‘Chief Librarian’ as he considered ‘Chief Librarian’ to be an outlier. However, he did not mention which job titles were excluded in his write up. He took a screenshot of the total number of ‘librarian’ employees per year from the CoV’s open data portal without indicating what the number means. He did not cite the used data source (as shown in figure 4.14).
Since participants were required to write the findings within a short period of time, and the task instructions did not mention explicitly what should be written in the reply, they might have lacked enough time and motivation to engage in deeper thinking on how the findings should be presented. Even considering the time constraints, there were many critical errors made in presenting the data, such as erroneous and unlabeled charts, misleading conclusions, and limited or no reference to data sources.

4.2.3 Summary of obstacles faced

The study finds that participants faced several types of obstacles, including lack of familiarity with the CoV’s open data portal, design limitations of the CoV’s open data portal, time constraints, and data literacy gaps. Since all participants used the CoV’s open data portal for
the first time in this study, participants needed to take time to familiarize themselves with the system, such as figuring out how to search for datasets and how to use the ‘Analyze’ tool to visualize the dataset. Time constraints clearly influenced their behavior and performance:

“I feel I am kind of rushing, as I only have 30 minutes. It usually takes a longer time to, like browse through and get familiar with the website, or try to find what information I need. I feel like I am kind of rushing.” (N02, SA-LDLG).

“I think people will perform differently when they are under a time constraint. When you are under a time constraint, you know one specific thing you need to find. Sometimes, you know it makes you more anxious to try to find that specific thing. Maybe you don’t browse the way you naturally do.” (N11, SA-HDLG).

The design limitations of the CoV’s open data portal also hindered participants to accomplish tasks. For example, participants struggled to use the ‘Analyze’ tool because there was no help or documentation embedded in the tool to guide participants on how to use it, and the high volume of information and options on the site made selecting an appropriate search tool challenging for some.

Obstacles faced by participants when selecting the most appropriate dataset, analyzing data, and presenting the analyzed findings seemed to stem primarily from the data literacy gaps, some of which participants were aware and others, not. Some participants did not know how to evaluate data and hence they selected inappropriate datasets, which could not answer the given problem. Participants were unaware of the need and importance of data wrangling and therefore they did not realize that their data was incomplete or inconsistent. Some participants did not know how to analyze quantitative data in a graph and created meaningless graphs. Finally,
participants did not know that they should cite the used data sources when presenting the analyzed data to someone else.

**4.3 Learning strategies**

Although participants faced a variety of obstacles to reaching successful task outcomes, they were most conscious of those related to system use, such as the CoV’s open data portal and Microsoft Excel. Participants were lacking a broad sense of how to carry out this data-focused task effectively, including components of data literacy. Key aspects of data evaluation, data wrangling, data presentation, and data citation were invisible to participants. Hence, they primarily tried to learn how to use different system components, rather than viewing their challenges through a data literacy lens.

When facing obstacles in using systems, participants took a consistent approach to overcome them. Many of them started with a trial-and-error approach. If the trial-and-error approach did not work as they expected, they actively sought step-by-step instructions. If they could not find written help, they asked for assistance from the researcher.

**4.3.1 Trial-and-error**

Trial-and-error is a problem solving method that takes the form of trying out different approaches and then eliminating the failures through observation until the task is completed successfully (Popper, 2002). Trial-and-error was the most employed learning strategy by participants in both SA-HDLG and SA-LDLG when they faced obstacles, especially when they faced obstacles due to the lack of familiarity with the systems.

As an example, participant N03 in the SA-HDLG randomly picked any employee with the title ‘Librarian II’ when she analyzed the salary for the first occupation. However, she unexpectedly found the data fluctuated across the past five years. She reflected on her first
experience and she planned to try a different way to analyze the salary for the second occupation. She chose to look for the salary of a selected employee named ‘P. Anderson’ rather than look for the salary of any employee with the title ‘Firefighter’.

In another example, participant N08 in the SA-LDLG first filtered only the Vancouver Public Library Board (VPLB) department in the ERE-75K dataset and then looked for the minimum and maximum remuneration for librarians by scrolling down the microdata. However, he found that there were many job titles other than ‘Librarian’ such as ‘Human Resource Consultant’ and ‘Security Coordinator’, and the remuneration data was not listed from smallest to the largest. He had to manually skip other job titles and looked for the lowest remuneration for a librarian when scrolling down the table. Thus, he reflected on his experience and tried to add another text filter as “librarian”. With only librarian data extracted in the table tab, he did not have to manually skip other job titles. Then he also applied a sort function on the remuneration column so that the remuneration data is listed from the smallest to the largest.

“94 thousand …81 thousand…86 thousand…81, 84, 85, 78, 80, 82…Ok, here is the lowest salary, 75502…um….um….oh, I think I can filter librarian…now I can sort [i.e. from smallest to the largest]. I find the range from 75502 to 97820” (N08, SA-LDLG)

His experience in learning how to use filter and sort function in the table tool of the CoV’s open data portal aligns with Kolb’s experiential learning cycle (1984). He started at the concrete experience stage by having an experience using the table tool in the CoV’s open data portal to find the range of librarian salary. Then, he entered the reflective observation stage to reflect on his experience. After reflection, he passed to the abstract conceptualization stage and he concluded that his method was able to find the range of librarian salary, but it was inefficient. At last, he planned to apply an additional filter and sort function at the active experimentation stage.
and tried out his plan. He was able to apply these strategies successfully when analyzing the salary of the second selected occupation.

In another example, participant N10 in the SA-LGLD wanted to analyze a subset of the ERE-75K dataset with only VPLB data, but he did not know which options he should choose to export the data. So, he tried out two different options to see which worked better for him.

“I am not sure if I download the whole dataset for every job or just the selected department…. I realize that this has every single department listed under the portal. So I am going to download the other option. Only the selected records, which probably will have the department I want. Indeed, it does.” (N10, SA-LDLG).

4.3.2 Step-by-step instructions

If participants could not overcome obstacles by trial-and-error, participants in both SA-HDLG and SA-LDLG actively searched for clues or step-by-step instructions online. For example, participant N01 in the SA-HDLG searched for “library” but no relevant datasets were retrieved, so she changed the search term to “librarians” and searched for datasets again. She reflected on her experience and tried out a different way, however, she still could not learn from her experience to overcome this obstacle. She needed external help to reflect on her experience. Thus, she went back to the study website to look for possible keywords. The study website provided clues for her to review her experience so that she could plan another approach to search for datasets.

“I am going back to the instructions and see what they need. 30 minutes to locate the needed data in the data portal. Employee salary. Ok, cool. Maybe if I narrow it down to employee salary it will work. Yes, ok.” (N01, SA-HDLG)
When participant N13 in the SA-HDLG could not find any datasets after trying several search queries such as “librarian salary” and “librarians”, she looked for instructions by browsing the “Help & API” page on the CoV’s open data portal.

“I am confused why nothing is coming up [i.e. after she tried two different search queries to search for datasets]. I am going to check on this page [i.e. “Help & API” on the menu bar of the CoV’s open data portal homepage] to see if any directions” (N13, SA-HDLG).

Participant N10 in the SA-LDLG also tried to locate a step-by-step guide from the “Getting started with open data” on the CoV’s open data portal homepage. However, it linked to the VPL’s open data research guides which gives a general guide on open data, rather than specific instructions:

“I am thinking about clicking onto this [i.e. “Getting started with open data” link], to learn more about the open data portal and how to use it. I am scrolling through this website [i.e. VPL’s open data research guides]…..I don’t think most of these bullet points are useful for me to search for salary of librarians. As these are only websites about open data, not sure if I can find any steps on how to find librarian salary” (N10, SA-LDLG).

When participant N05 in the SA-LDLG did not know how to apply a filter in Microsoft Excel, she searched for ‘how to make filter in excel’ in Google search engines to obtain step-by-step instructions to filter data in Microsoft Excel. Participant N10 in the SA-LDLG also searched for “how to find the lowest value on excel” in Google search engines to look for a formula to calculate the minimum value in Excel.

“I am going to do a function that allows me to choose the lowest value under the remuneration of a librarian. I am not sure what the function is. I am going to search on Google.” (N10, SA-LDLG).
Although the CoV’s open data portal provides complete help guides that have step-by-step instructions on searching datasets and creating charts on the “Help & API” page (as shown in figure 3.5), no participant was able to locate them. Instead of checking the complete help guide, participants clicked “Getting started with open data” which linked to VPL’s open data research guide. Unlike the complete help guides, the VPL’s open data research guide contains generic information about open data. Participant N09 in the SA-LDLG and participant N11 in the SA-HDLG checked the VPL’s open data research guide but they did not find it useful. They mentioned in the semi-structured interview that they would like to have step-by-step instructions on the CoV’s open data portal.

“For me, it would be great if they have more instructions, like on the portal itself. To have a little bit more instructions on how to get you to where, and what information you are looking for. Kind of like a FAQ page almost. Just some very generic like what to click on to plot data points” (N09, SA-LDLG).

“If I click here [i.e. “Getting started with open data”], I would expect to see a video tutorial on how to get started. Maybe it would be good to have a very quick introductory tutorial like how to use our portal for a beginner, how to navigate our dataset.” (N11, SA-LDLG).

Participant N10 in the SA-LDLG and participant N11 in the SA-HDLG mentioned in the semi-structured interview that they would like to have some guidance at the time when they need it rather than searching for guidance from another page.

“Maybe there is a help button. The format will be like what does this button does on the ‘Analyze’ page. For example, if you press down the scroll list, what is this scroll list for,
what is this graph tool for. For like each specific function on the ‘Analyze’ page...I appreciate it if there are instructions on how to use it properly.” (N10, SA-LDLG).

“I hope there is an information bar to tell you what everything means [i.e. functions on the analyze tab in the CoV’s open data portal]. I don’t know if this meant to be used by data scientists. If it is the case, I feel like the breakdown series is very obvious what it means. But I think maybe for the average users, what does break down series mean?” (N11, SA-HDLG).

4.3.3 Asking for advice

Asking for assistance was another option if participants could not overcome obstacles by trial-and-error and/or participants could not find step-by-step instructions. For example, after participant N06 tried many ways but still could not find relevant datasets from the CoV’s open data portal, he asked the researcher what he can do. When participant N06 in the SA-HDLG was looking for data from the dataset schema and was unable to find data for the selected job title, he asked the researcher to guide him on how to obtain this data.

“There is a lot of job title here, but as long as I click one title. For example, like this one [i.e. Aquatic 1]. It doesn’t update [i.e. dataset schema]. It is still the administrative manager 3. So what can I do to update the information?” (N06, SA-HDLG).

When participant N09 in the SA-LDLG could not find 2018 data from the WPRG dataset, she asked the researcher whether she was on the right page for the study. Once she located the ERE-75K dataset, she asked the researcher again whether she was on the right page. These participants made use of social interactions as a learning strategy to seek help from others to stimulate their learning.
To address the third research question of “what, if any, learning strategies do users employ when they meet obstacles in using open data to accomplish a task?” the study finds that trial-and-error was a common learning strategy that participants employed when they faced obstacles. Some participants also looked for step-by-step instructions to overcome obstacles, however, in his study they were seldom able to access these within the CoV’s open data portal. If the participants could not overcome obstacles by themselves, some of them would ask someone for advice and guidance to help them, although this was not supported by the portal itself, but only possible due to the conditions of the user study.

4.4 Summary of research findings

Participants in both SA-HDLG and SA-LDLG primarily used the CoV’s open data portal to make use of open data to accomplish the assigned tasks. External tools, such as Microsoft Excel, Microsoft Word, Apple Pages, and Snipping Tool, were used to a less extent. The participants in the SA-HDLG used external tools as supplementary options when data could not be analyzed within the CoV’s open data portal. Some participants in the SA-LDLG primarily used the external tool, Microsoft Excel, to accomplish tasks while others carried out the tasks within the CoV’s open data portal.

Study participants faced obstacles in figuring out the relevant pages to perform tasks, formulating search queries to search for datasets, selecting the most appropriate datasets among all the retrieved datasets, wrangling data, choosing the appropriate data analysis methods, using the data analysis tools to analyze data, presenting data properly for the target audience, and citing the used data sources. Participants in the SA-HDLG mainly faced the obstacles when using the data analysis tools in the CoV’s open data portal. Participants in the SA-LDLG faced obstacles when processing datasets and using the data analysis in the CoV’s open data portal. They noticed
that they faced obstacles when using systems like the ‘Analyze’ tool in the CoV’s open data portal and Microsoft Excel. However, they were often unaware of mistakes or important gaps in their process when they were evaluating, processing, and analyzing data, as they seemed to be unaware of their limited knowledge. This is a more challenging issue to address.

When they faced obstacles that halted their progress, participants took a trial-and-error approach to overcome them. If they could not use the trial-and-error approach to overcome the obstacles by themselves, some participants tried to look for step-by-step instructions or asked someone (the researcher, in this case) for advice. There were no apparent differences in the learning strategies employed by participants in the SA-HDLG and the SA-LDLG.

Overall, results suggest that even a well-designed open data portal, such as the one used in this study, can be challenging to use for first time users. Even though participants in this study, all enrolled in post-secondary programs, have higher educational levels than average citizens, they still faced many obstacles when completing the given tasks. Some participants, mainly in the SA-LDLG, were unable to overcome recognized obstacles such as formulating appropriate search queries and using the ‘Analyze’ tool on their own, and they need some external help. Some obstacles were unrecognized by participants in both SA-HDLG and SA-LDLG, such as the need to wrangle data and the importance of including key information when presenting data for the target audience, which caused participants to make mistakes when using open data.

Participants with lower data literacy also overestimated their ability to process data in the self-assessment, which suggests that people with insufficient data literacy skills are at a higher risk to use open data incorrectly, without awareness of those errors. The results clearly indicate that having an open data portal with accessible datasets and data processing tools does not mean people can make use of open data properly on their own. People also need data literacy skills to
use open data. Since the majority of participants spent a long period of time on the open data portal to accomplish tasks, it creates an opportunity to design learning tools embedded within the open data portal to help users to develop data literacy skills so that people can use open data effectively.
Chapter 5: Discussion

5.1 Key insights of research findings

Open data portals, such as the CoV’s open data portal, are intended for everyone (City of Vancouver, 2019c). Hence, it is essential to design the open data portal for both beginner users and experienced users. This study investigated how participants with varying levels of data literacy interacted with open data and associated tools to make use of data, the obstacles they faced during the process, and the learning strategies they took to overcome the obstacles. The general results indicate that this group of well-educated young adults faced numerous challenges in completing simple tasks using a state-of-the-art municipal open data portal. These results support prior research indicating that low data literacy skills constitute a barrier to use of open data (Gurstein, 2011; Open Data Barometer, 2017) and provide further evidence of the need for including guidance (Janssen et al., 2012), and data literacy training (World Wide Web Foundation, 2018) to support inexperienced users to make use of open data in order to make the benefits of open data widely available.

The research findings reveal some insights for designing open data portals that can help both beginners and more experienced users learn to make effective use of open data.

5.1.1 Profile of low and high data literacy users

It is challenging to assess an individual’s data literacy skills. In this study, a self-assessment approach was adopted to classify participants into two groups, the low data literacy group (SA-LDLG) and the high data literacy group (SA-LDLG). However, the overall high self-assessment scores made it difficult to identify two groups based on a simple median split, such that we needed to rely on a subset of questions to identify high and low groups. More work is needed to refine the instrument developed for this study. Further, there are some discrepancies between the self-
assessment results and the participant’s actual performance. For example, participant N06 performed poorly relative to other participants, although he rated himself highly and he was grouped into the SA-HDLG. The discrepancies and generally high scores are likely due to the Dunning-Kruger effect, in which poor performers are unaware of their incompetence and hence they overestimate their abilities (Dunning, 2011). Despite some discrepancies in the self-assessment results of participants’ data literacy skills, the study can still reveal differences and similarities among participants with different levels of data literacy skills. For example, the interactions and obstacles faced by the majority of participants in the same group were similar and there were some apparent differences between the general behaviours of the two groups, such as the length of time and effort expended to find the dataset was higher for the lower data literacy group.

Participants with higher data literacy skills appeared to be more willing to learn to use the open data portal, although some of them found this difficult. Participants in the SA-HDLG primarily used the CoV’s open data portal to find and use the dataset. However, participants with lower data literacy skills were more likely to switch to external tools when they faced obstacles to using the open data portal. Many participants in the SA-LDLG primarily used the CoV’s open data portal to find the dataset and then used Microsoft Excel to analyze the dataset. Even though they also faced obstacles to using Microsoft Excel, they preferred to use the more familiar external tool rather than learn to use the built-in tools in the portal. The obstacles faced by participants with higher data literacy skills were mainly related to using unfamiliar data analysis tools. Many of them were able to overcome obstacles without assistance. Participants with lower data literacy skills faced more obstacles than participants with higher data literacy skills as they found it hard to learn to use the CoV’s open data portal, and also struggled to find data, handle data, and to use
external data analysis tools (i.e. Microsoft Excel). They showed less understanding of what to do when they obtained the dataset and tended to need external help to overcome obstacles. These results suggest that users with low data literacy skills may need more higher level conceptual tools about the nature of data and approaches to analysis, including step-by-step guidance to support users to build up data literacy; while those with higher levels of data literacy may benefit more from improving the usability of the tools to support awareness of and use of system features to perform tasks.

5.1.2 Obstacles

Although the study tasks focused on competencies related to finding and using data, it was clear that participants needed all five core competencies in the Calzada Prado & Marzal’s data literacy competency framework to complete the tasks successfully. These five core competencies include (1) understanding data, (2) finding and/or obtaining data, (3) reading, interpreting and evaluating data; (4) managing data, and (5) using data (Calzada Prado & Marzal, 2013).

In this study, Task 1 required participants to find the needed dataset. To complete Task 1, participants needed to understand how datasets are structured and managed in the CoV’s open data portal. Unlike popular search engines like Google which index all terms in a document, the CoV’s open data portal only indexes the metadata of a dataset but not all the available terms in the dataset. Some participants could not retrieve any dataset because they expected full-text searching in the CoV’s open data portal and so they used search queries such as ‘librarian’ and ‘librarians’. Lacking the ‘understanding data’ and ‘managing data’ core competencies, participants faced challenges to retrieve relevant datasets in open data portals which usually describe datasets in metadata rather than indexing all terms in the datasets (Máchová et al., 2018). Some participants also faced obstacles to figure out whether to search datasets by
keyword searching or by browsing. Improving the usability of the CoV’s open data portal by keeping the interface simple (Hearst, 2009) may help users to overcome this obstacle.

After retrieving relevant datasets, participants were required to first evaluate all the possible datasets and then select the most appropriate one to complete Task 1. These two steps are related to ‘finding and/or obtaining data’ and ‘reading, interpreting and evaluating data’ core competencies. According to Calzada Prado & Marzal’s (2013), individuals with ‘finding and/or obtaining data’ core competency need to be aware of possible data sources, detect whether data sources can solve a given problem, and select the most appropriate data source to the given problem. Some participants successfully obtained the needed dataset within a few minutes because they evaluated whether the possible datasets contain the needed data by exploring the microdata in the ‘Table’ tool. If the microdata did not contain the needed data, they checked out the other possible datasets. Other participants found this challenging (N05, N09, and N12 in the SA-LDLG) or failed (N06 in the SA-HDLG) to locate the needed dataset because they were unaware of a possible dataset (i.e. the ERE-75K dataset) and they did not evaluate the microdata of the selected dataset. If individuals do not understand how datasets are typically structured and described and are not aware of the features of a dataset that can be used to assess relevance and quality, they are less likely to find appropriate datasets to meet their information needs.

Once the most appropriate dataset is selected, an individual should first understand the data before analyzing it. The ‘understanding data’ core competency indicates that a data literate person needs to know the data definition, data types, and how data was generated and by whom (Calzada Prado & Marzal, 2013). In this study, the needed dataset (i.e. ERE-75K dataset) was generated by the Government of the City of Vancouver every year. Since the organizational structure of government may change over time, the job titles and names of departments may not
be consistent across different years. Therefore, data wrangling is necessary before analyzing the dataset. However, none of the participants addressed data errors or inconsistencies, which affected the accuracy and/or completeness of their data analysis. When an individual has a limited understanding of data, they are more likely to misuse the data and knowingly or unknowingly produce inaccurate results.

In this study, Task 2 required participants to analyze the needed dataset. Participants need to have knowledge of data analysis methods and data analysis tools. Many participants in the SA-HDLG demonstrated basic knowledge of data analysis methods as they knew how to summarize data in a table or a graph. They mainly faced obstacles in figuring out how to use the ‘Analyze’ tool in the CoV’s open data portal. Participants in the SA-LDLG also found it hard to use the ‘Analyze’ tool in the CoV’s open data portal. For example, participant N11 in the SA-HDLG could not figure out how to unstack a column chart (as shown in figure 4.7) while participant N02 in the SA-LDLG did not know how to filter job titles in the column chart (as shown in figure 4.9). Systems with high usability should provide visible instructions for use of the system or easily retrievable help and documentation focusing on the user’s task in order to support users to use the systems effectively (Nielsen, 1994). The low usability of the ‘Analyze’ tool in the CoV’s open data portal, resulting from hidden functionality and lack of help and documentation, hindered participants to use data analysis tools effectively.

Some participants in SA-LDLG (N07 and N09) not only faced obstacles in using the data analysis tools but also encountered challenges in choosing the correct methods to process data. Hence, they created meaningless graphs (as shown in figure 4.11 and 4.12). Improving the usability of an open data portal can help users to learn to use the data analysis tools in the open data portal, however, it may not be able to help users to learn data analysis methods. This
suggests the need for support and guidance on appropriate methods of analysis and user feedback when incorrect methods are used or questionable results are produced.

In this study, Task 3 required participants to present the analyzed findings in a write-up to a friend. When presenting analyzed findings, a data-literate person should make ethical use of data by acknowledging the data source, and present the used methods and interpreted results transparently and honestly (Calzada Prado & Marzal, 2013). Few participants in the SA-HDLG and none of the participants in the SA-LDLG cited the data source or identified the data analysis methods used. This issue seems to stem primarily from a lack of awareness of the importance of providing this information, and suggests that an open data portal should provide additional features to support data sharing and referencing so that users can learn through reflective observation. For example, the portal may build in a sharing function that by-default extracts the data processing steps and data citation together with the analyzed findings.

5.1.3 Learning Strategies

In this study, many participants, regardless of their data literacy levels, learned to use the system such as the ‘Table’ tool and ‘Analyze’ tool in the CoV’s open data portal using a trial-and-error approach. They started by trying out a few functions in the tool and observed what happened. If they could not obtain the results as expected, they tried other functions until they were satisfied with the outputs of the tool. Heavy use of this approach is not surprising, as trial-and-error is a commonly used approach to learn informally in the workplace (Jeong, McLean, & Park, 2018; Marsick & Volpe, 1999). The trial-and-error learning strategy aligns with the Kolb’s informal learning cycle (Kolb, 1984) offers some opportunities for support. Many participants started at the concrete experience stage to work on a task using a selected tool. Then they observed the outcome and reflected on their experience at the reflective observation stage. After reflection, they
concluded what they learned at the abstract conceptualization stage. If they did not succeed at the task, they drew upon the lessons learned at the abstract conceptualization stage and planned for a second attempt. For example, participant N01 in the SA-HDLG first changed the value in the ‘Break down series’ function to ‘Remuneration’ in the ‘Analyze’ tool, then she observed the new graph. The new graph was meaningless, so she removed the ‘Remuneration’ value in the ‘Break down series’ function. Based on her first attempt, she learned that the ‘Break down series’ function was not useful to her. Therefore, she tried out the ‘Y-axis’ function in her second attempt. She changed the value in the ‘Y-axis’ to ‘Average remuneration’ and then she created a graph that she wanted.

The trial-and-error approach was more effective among participants with higher data literacy skills than those with lower data literacy skills. Some participants with lower data literacy skills were not able to identify the mistakes they made, and therefore did not learn from them. After the first attempt in a trial-and-error approach, they might conclude that the task was successful at the abstract conceptualization stage even though they made a mistake. Since they did not notice that they failed the task, they did not try out a second approach to work on the task. As a result, they learned an incorrect approach to the task. For example, similar to participant N01 in the SA-HDLG, participant N09 in the SA-LDLG also changed the value in the ‘Break down series’ function to ‘Remuneration’ in the ‘Analyze’ tool. However, she did not notice that the resulting graph was meaningless. Therefore, she used the meaningless graph to complete the tasks (as shown in figure 4.12). In an informal learning environment, learners with insufficient skills have a higher risk to reach the wrong solutions as there is limited guidance to help them to carry out tasks correctly (Marsick & Watkins, 2018). Even if participants recognized the mistakes they made, participants with lower data literacy skills had limited knowledge to correct their mistakes on their own.
Participants in the SA-HDLG were able to learn informally by themselves because they were able to reflect on their prior knowledge. Participants N11 in the SA-HDLG mentioned in the interview that she had analyzed datasets using other data analysis tools, such as Microsoft Excel and Tableau, before so she was able to rely on her experience to figure out how to use the ‘Analyze’ tool in the CoV’s open data portal. In contrast, many participants in the SA-LDLG, such as participant N02 and N10, could not figure out a way to overcome obstacles by themselves. They tried to seek external help such as looking for step-by-step instructions and ask someone for advice. However, due to limited external help in the CoV’s open data portal, they failed to correct their mistakes by the trial-and-error approach.

Since many participants recognized that they faced obstacles to using the system, they employed a trial-and-error approach to informally learn using the system. Apart from learning to use the system, participants needed to learn other aspects of data literacy skills such as understanding data and making ethical use of data. Other essential data handling activities, such as data evaluation, data wrangling, and data citation, were invisible to participants due to the lack of knowledge. Hence, they did not know that they need to do the tasks and missed the opportunity to learn by doing them. Results of this study indicate that the trial-and-error approach, which is the most common approach for users to learn how to use unfamiliar systems, is not effective as a means of learning broader conceptual data literacy skills. Therefore, open data portals should provide novice users with the opportunity to develop their data literacy competencies while engaging with the portal.

5.1.4 Opportunities for supporting learning

Even if an open data portal provides comprehensive data processing functions, users will miss the learning opportunities to use open data if they leave the portal due to low usability.
Nielsen (2012) argued that poor system usability not only adversely affects users to use the available features of the website, but it also causes users to leave the site. This study found that participants with lower data literacy skills were more likely to leave the open data portal than participants with higher data literacy skills. Therefore, having an open data portal with high usability is the foundation to design a portal that can support users to learn data literacy skills informally. Once the usability of an open data portal is optimized, various learning tools can be employed to encourage and support users to develop data literacy skills across the stages of the informal learning cycle (Kolb, 1984).

In this study, many participants learned to use the CoV’s open data portal by a trial-and-error approach, which aligns with the Kolb’s learning cycle (Kolb, 1984). Taking learning to use the ‘Analyze’ tool using a trial-and-error approach as an example, participants passed through Kolb’s learning cycle by carrying out different activities. Some possible tools to support participants at each stage of the learning cycle are shown in Figure 5.1 and discussed in more detail below.
Figure 5.1 Kolb’s learning cycle in the context of using the ‘Analyze’ tool in the CoV’s open data portal.
Adapted from Kolb, 1984.

Concrete Experience
- e.g. Used a function in the 'Analyze' tool
- (Possible support: Step-by-step wizards)

Active Experimentation
- e.g. Planned to a function in the 'Analyze' tool
- (Possible support: FAQ)

Reflective Observation
- e.g. Reviewed how the graph was changed
- (Possible support: Discussion forum)

Abstract Conceptualization
- e.g. Learned how the function changed the graph
- (Possible support: Video tutorials)

The CoV’s open data portal provides a variety of tools, such as the search system and built-in data analysis tools for users to find and use open data within the portal. Many participants used these tools and started learning at the concrete experience stage. Participants in the SA-HDLG had a clear idea of what steps should be taken to find and use the dataset due to their prior knowledge. Learning to use unfamiliar tools is the major challenge they faced. Some possible supports to help experienced data users at the concrete experience stage can be tooltips to increase awareness and understanding of the available tools and directly accessible documentation to explain the functions of data tools and help users to reflect on the experience in light of their prior knowledge. Participant N11 in the SA-HDLG mentioned in the interview that the ‘Analyze’ tool can be improved by providing information explaining how each function of
the ‘Analyze’ tool works. This finding aligns with a usability study of the Dublin City’s open data portal, in which many study participants who were experienced data users commented on the problem of the lack of tooltips and pop-up notes on the open data portal to guide use of a built-in data analysis tool (Osagie et al., 2017). Participants in the SA-LDLG faced different challenges in using the tools and using the dataset at the concrete experience stage, as they had less prior knowledge to draw upon. Due to limited knowledge of how to handle data, tooltips that explain system use may not be enough to help inexperienced data users. If the user has limited skills or knowledge, he/she will not know he/she did something incorrectly (Dunning, 2011).

Watkins et al. (2018) found that informal learners who have insufficient skill and knowledge may carry out tasks incorrectly and reach the wrong solutions because there is limited guidance to help them to reach the correct solutions. Therefore, inexperienced data users need additional support at the concrete experience stage. To help inexperienced data users to develop data literacy skills and to work on tasks correctly, some possible learning tools could be walk-through tutorials or step-by-step wizards that contain a sequence of dialog boxes that guide users to go through steps in a well-defined order. For example, a walkthrough tutorial can guide users to search for datasets in a sample case; a step-by-step wizard can guide users to analyze a dataset and produce a graph, and users can follow the steps to choose the type of graph and set dimensions for x-axis and y-axis. Under the guidance of a walk-through tutorial or step-by-step wizard, users can learn to process data at the concrete experience stage. These types of tools have the advantage of providing more contextual and explanatory information to build understanding.

After doing tasks at the concrete experience stage, informal learners reflect on the experience at the reflective observation stage (Kolb, 1984). However, there is very limited support in the CoV’s open data portal to help users to reflect on the experience. Some
participants who were experienced in using data, mainly in the SA-HDLG, reflected on their own experience. For example, participants \textit{N11} in the SA-HDLG mentioned in the interview that she previously analyzed datasets using other data analysis tools such as Microsoft Excel and Tableau, so that she can rely on her experience to figure out the way to use the ‘Analyze’ tool in the CoV’s open data portal. Some participants who were inexperienced in using data, mainly in the SA-LDLG, were looking for step-by-step instructions or asking someone for advice to help them reflect on the experience to overcome the obstacles. For example, participant \textit{N02} in the SA-LDLG could not figure out how to filter data by selected job titles and he received an error message “\textit{There are too many series to be displayed correctly, try to refine your query a bit}” popped up (as shown in Figure 4.9). Since he could not figure out the way to refine the query, he sought steps to refine the query, but step-by-step instructions were not available. Beginners who have little idea how to use data may benefit from support from experienced users. Some possible learning tools to help them can be a chatroom and/or a discussion forum. Both a chatroom and discussion forum provide a platform for users to ask questions to other members of the community. In this study, participants demonstrated their interest in such a feature by asking for advice from the researcher, something which would not normally be available to them in using the system on their own. Discussion forums could provide a channel for these participants to get support from a wider community. Discussion forums are also common features in many data visualization systems such as Tableau, Microsoft Power BI, and Qlik (Microsoft, n.d.; Qlik, n.d.-c; Tableau, n.d.). Osagie et al. (2017) found that users do not only want to search, analyze and use datasets in an open data portal, they also want to engage with other members of the community to discuss issues related to the data. These types of learning tools could allow users to seek external help to learn at the reflective observation stage.
In this study, some participants had data literacy gaps, such that they were not aware of the need for data evaluation, data wrangling, and data citation. Therefore, they did not do these tasks. In other words, they could not develop these data literacy skills through concrete experimentation or reflection because they did not engage in them. To fill this knowledge gap, informal learners may need to learn by starting at the abstract conceptualization stage or the active experimentation stage. A possible learning tool to help users to learn at the abstract conceptualization stage can be a data literacy online e-learning program to teach users how to make use of data. The CoV’s open data portal provides the “European Data Portal eLearning program”, which is a comprehensive e-learning program teaching users about the general concepts of open data and the use of open data (“European data portal eLearning program,” n.d.). However, it is unknown whether the e-learning program is a useful tool to help users to develop data literacy skills as the time-limited nature of this study meant that participants could not engage in this type of learning. Providing frequently asked questions (FAQ) or video tutorials may be a lighter-weight means of helping users to learn at the active experimentation stage, without taking an entire e-learning course. For example, a ‘how to choose a dataset’ question with step-by-step instructions as the answer could be provided in the FAQ section so that novice users could learn how to evaluate and select the dataset before doing the task. In another example, a video tutorial can show step-by-step instructions on how to select the appropriate methods to analyze datasets based on a sample problem and the characteristics of data variables. After learning the steps for processing data through the FAQ section or a video tutorial, users can move from the active experimentation stage to try out the steps on their own at the concrete experience stage. Both videos and FAQs were mentioned by participants in the study, as tools they had hoped to find to assist them.
The learning tools should focus on supporting users at different stages of the informal learning cycle. For example, a FAQ section and video tutorial may help users to learn at the active experimentation stage; a walk-through tutorial may guide users to learn at the concrete experience stage; a chatroom or discussion forum may allow users to ask for guidance at the reflective observation stage; an online e-learning program may help users to learn a broad concept of data processing at the abstract conceptualization stage. These recommendations align with Máchová et al. usability evaluation of open data portal and Zhu et al. user engagement framework that providing support to users to use the open data portal is one of the key criteria of a useful open data portal (Máchová et al., 2018; Zhu & Freeman, 2019). Máchová et al. (2018) suggest supporting users through high quality of documentation and tutorials, discussion forums, user rating and comments, and social media feedback and sharing. Zhu et al. (2019) suggest supporting users through FAQs, showcase/examples, and community discussion. These prior studies did not go further in investigating how these supporting tools are beneficial to users with different levels of data literacy skills. This study found that users with lower data literacy skills need support on learning the broad concepts of data literacy while users with higher data literacy skills need support on using unfamiliar data processing tools. Since users with different data expertise levels need different support, it may be better to design different interfaces with learning tools for beginners and advanced users rather than providing a single interface with universal learning tools for all users. The interface for beginners may focus on designing step-by-step wizards and showcase examples to guide users to open data while the interface for advanced users may focus on designing tooltips and quick tutorials to guide users to use the built-in data processing tools within the portal. This study provides empirical evidence upon which to base concrete recommendations as well as
a theoretical framing from learning theory and data literacy research to inform the design of such learning supports.

5.2 Challenges and limitations

This study was exploratory and subject to the limitations of controlled user studies. The 30-minute time limit for the assigned tasks was set in the study to avoid overloading participants with a lengthy task; however, the time limit created pressure on participants, which influenced their performance. A few participants mentioned in the interview that one of the challenges they faced was the time limit. If there was no time limit, they may have performed better and they may have employed a wider range of learning strategies to overcome the obstacles they faced. For example, they may have spent more time exploring the CoV’s open data portal; reviewing the “European Data Portal eLearning program” to learn about open data, or may have asked more knowledgeable family or friends for help during the study. At the same time, the time limit helped to highlight differences between the high and low data literacy groups by forcing participants to rely upon their existing knowledge rather than learning while doing.

The self-assessment data literacy questionnaire was devised for this study by drawing upon an existing data framework. However, the instrument was not previously tested or validated, and it was difficult to clearly divide the participants into two groups using the resulting scores. While there did seem to be differences in behavior and performance between the high/low group, observing the performance of participants indicates that some self-assessments were higher than warranted. We suspect that the Dunning–Kruger effect, in which incompetent people have little insight into their incompetence and hence will overestimate their performance, came into play (Dunning, 2011). Since the results of the self-assessment data literacy questionnaire could not truly reflect their actual performance on handling data, the participants
were roughly grouped into two levels of data literacy, the higher data literacy group and the lower data literacy group, rather than ranking participants with different self-assessment scores. Participants’ performance and their think aloud commentaries were taken into account to reflect their interactions, obstacles, and learning strategies. Objective assessment of data literacy skills of participants such as evaluating their performance should be considered in the future study. Further, robust tools for assessment of data literacy are needed in order to conduct further research in this area.

Participants were university students who were young and well-educated, which does not reflect the general public and their use of open data. Moreover, the study made use of the CoV’s open data portal, the findings of this study may not apply to other open data portals. Future research should focus on more diverse users with a wider range of data literacy skills, and different open data portals.
Chapter 6: Conclusion

This study set out to understand how users with differing levels of data literacy use an open data portal, what obstacles they encounter, and what learning strategies they employ during use. This was accomplished through an observational study of 14 participants completing a series of simple data tasks using a feature-rich municipal open data portal. Study findings point to challenges arising from data literacy gaps and system usability issues, which are not currently addressed in this relatively sophisticated open data portal. Participants in the low data literacy group faced more obstacles due to gaps in understanding how to process open data than those in the higher data literacy group, and hence they need extra support to develop data literacy skills to make use of open data. Participants in the high data literacy group mainly faced obstacles in figuring out how to use the unfamiliar data analysis tools in the portal, which can be reduced by improving the system usability. Some important data processing activities such as evaluating data quality, data wrangling, labeling analyzed findings and data citation, seemed to be lacking across all participants, regardless of their data literacy self-assessment. The portal needs to provide additional features and learning tools to guide users to perform these key data activities.

Many participants employed a trial-and-error approach to accomplish tasks and at the same time, they learned new things such as, how to use the data analysis tools. The trial-and-error approach aligns with Kolb’s informal learning cycle. The trial-and-error approach was more effective to learn how to use the system rather than to gain a broader knowledge of how to process data. Participants with higher data literacy skills had basic knowledge on how to process data, so the trial-and-error approach was effective for them to learn to use the unfamiliar tools in the CoV’s open data portal. Participants with low data literacy skills faced challenging to accomplish tasks accurately using the trial-and-error approach as they either failed to identify the
mistakes they made due to the lack of knowledge or they did not have adequate skills to figure out a way to overcome obstacles they faced. When participants recognized their mistakes but failed to fix them on their own, they sought external help by looking for step-by-step instructions or asking someone for guidance. The limited support for system use and learning within the portal aggravated existing obstacles. The study reveals opportunities for designing more supportive learning environments to make open data portals more useful for a wide audience. Findings suggest that users with different data literacy levels need different learning tools to support them to use open data and develop data literacy skills informally. Recommendations are presented in the context of Kolb’s informal learning cycle, as users engaged in activities such as concrete experimentation and reflection while using the portal. Users with high data literacy skills may need tooltips or directly accessible documentation to teach them how to use the system. Users with low data literacy skills may need a variety of learning tools to guide them on how to process open data. Some possible learning tools include walk-through tutorials, step-by-step wizards, video tutorials, discussion forums, FAQs, and online e-learning programs to support users at each stage of the informal learning cycle. Future studies may explore how learning tools should be designed for open data portals that can not only help users with different levels of data literacy skills use the system, but also develop stronger data literacy skills.

This study showed that the majority of participants preferred to make use of data within the open data portal, but that is dependent upon the availability of comprehensive and user-friendly functionalities. To meet the goal of transforming an open data portal into a learning environment, it is important to keep them in the portal while working on tasks so that they may learn informally in the context of their own life needs and activities. Hence, prior to developing learning tools to support users to learn data literacy skills informally, open data portals must be
designed with high utility and high usability. Since all citizens should have the opportunities to make effective use of open data for citizen participation, governments should help citizens with low data literacy skills to develop the needed skills to use open data. An open data portal as a platform for citizens to access and use open data should provide learning tools to help citizens to develop their data literacy.
Bibliography


City of Vancouver. (n.d.). City of Vancouver’s Open Data Portal. Retrieved from
https://opendata.vancouver.ca/pages/home/

City of Vancouver. (2019a). Help, API, and licence. Retrieved from
https://opendata.vancouver.ca/pages/help/

City of Vancouver. (2019b). Open data portal. Retrieved August 17, 2019, from
https://opendata.vancouver.ca/pages/home/

City of Vancouver. (2019c). Welcome to the portal. Retrieved from
https://opendata.vancouver.ca/pages/portal-news/


Data To the People. (n.d.). Databilities: A data literacy competency framework. Retrieved from
https://92daafc6-7a9d-43e8-a3d6-37bb86881647.filesusr.com/ugd/1ff4ae_14805e0c8ef14b54bdafd38e44d5de23.pdf

Data To The People. (2020). Data To The People. Retrieved from
https://www.datatothepeople.org/


https://doi.org/10.1016/B978-0-12-385522-0.00005-6

European data portal eLearning program. (n.d.). Retrieved from
https://www.europeandataportal.eu/elearning/en/#/id/co-01

Executive Office of the President of The United States. (2009). President’s memorandum on


Socrata. (n.d.). City of Edmonton: Open data strategy. Retrieved from


https://doi.org/10.1016/j.giq.2016.02.001


Appendices

Appendix A User study documents

A.1 Study instructions of think-aloud observational session

Instructions
Please read the instructions carefully and ask questions, if any, before you proceed.

Situation:
Imagine that you have a friend who is graduating from his/her undergraduate study soon. He/she is asking for your advice on his/her future career. You know that he/she may be interested in being a librarian or another occupation. You think it will help to let him/her know how much these two occupations earn. The City of Vancouver releases data on employee salary (also known as remuneration or earnings) in their open data portal, and you think it may be a good source to get information. You will reply to your friend with your findings in a table(s) or graph(s).

Work Tasks:
You will have 30 minutes to complete the following work tasks:

7. Locate the needed dataset in the City of Vancouver’s open data portal
8. Make use of located dataset to analyze the last 5 years salary (also known as remuneration or earnings) of
   a. librarians and;
   b. one occupation that your friend may be interested in (e.g. planner, analyst, etc.)
9. Write your reply using the reply template
You may use any computer applications available on your desktop computer/laptop if you find these tools are helpful to complete the tasks.

**Think-aloud Instruction:**

When you are working on the tasks, I would like to say your thoughts out loud that:

1. What are you thinking;
2. What are you about to do, and why;
3. What works well for you; and
4. What is difficult?

The researcher will prompt you or ask you questions to share your thoughts when you are working on the tasks.

**Reply Template**

**Instructions:**

1. Please copy the below reply template to any word processing software (e.g. Microsoft Word, Google Docs, Apple Pages).

2. Create your content by editing text in blue color according to the instructions

*I am glad to hear that you are graduating soon. Refer to your previous message, I find some information about the salary of two occupations that you may be interested in.*

*Librarian*

<Insert a summary of the findings related to librarians’ salary in the last 5 years. Interpret the findings in table(s) or graph(s).>

<Add any relevant / important information>

<Insert one occupation you selected here>
<Insert a summary of the findings related to one occupation's salary in the last 5 years.>

*Interpret the findings in table(s) or graph(s).*

<Add any relevant / important information>
A.2  Self-completion questionnaire

Using Open Government Data (Online Study)

Participant Number:

What is your current program level?
- Undergraduate
- Master
- PhD
- Other, please specify:

Which faculty/school are you in?
- Faculty of Applied Science
- Faculty of Arts
- Faculty of Dentistry
- Faculty of Education
- Faculty of Forestry
- Faculty of Land and Food Systems
- Faculty of Medicine
- Faculty of Pharmaceutical Sciences
- Faculty of Sciences
- Peter A. Allard School of Law
- Sauder School of Business
- Other, please specify:
What is your age?

- 17-21
- 22-26
- 27-31
- 32-36
- 37-41
- 42-46
- 47-51
- 52 or above

What level of knowledge do you have about open government data?

- No knowledge at all
- Very little knowledgeable about it
- Little knowledgeable about it
- Some knowledgeable about it
- Very knowledgeable about it

Have you used any open government data website before?

- Yes
- No
Have you used the new City of Vancouver open data portal (i.e. released in August 2019)?

- Yes
- No

For each statement, please use the following scale to indicate your level of agreement:

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I can identify different kinds/types of data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can describe how data is produced in society</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can describe a wide range of possible applications of data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can find credible data to solve a given problem or need.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can evaluate the quality of data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can identify the limitations of data for a given problem or need.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can describe the various ways to represent data (e.g. written, numerical or graphic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can interpret data in the form of charts or tables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can describe data with metadata</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can manage data for subsequent reuse</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can clean and prepare data for analysis</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can analyze data using tools such as spreadsheets or statistical software</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can summarize the results of data analysis using various forms of representation (e.g. written, numerical or graphic)</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can use data ethically (e.g. protect confidentiality, data license)</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can cite data sources I have used</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A.3  Researcher’s script during the study session

Introduction & Technical Setup

Hello. My name is Winnie Li. I am a Master student at the University of British Columbia studying Library and Information Studies. I will be running the study sessions with you today.

First, we need to do the technical set up. Can you please share your screen with me? Remember to close any applications that may contain sensitive information before you share your screen with me. Please share the whole screen, it should be the first option of screen sharing.

[Wait for the participant to share screen]

Let’s test out if the screen sharing works well. May you please try to open an application like MS Word.

[Wait for the participant to open MS Word or another application]

I just send you a link of the study session via the chatroom. Please open the link in any web browser. (Note: send out link https://blogs.ubc.ca/ogdresearch/)

[Wait for the participant to open the study website]

On the homepage, you can see the agenda of today session.

Consent

I would like to start by making sure you understand what we will do in the session and that you consent to participate. Did you have a chance to read the consent form I sent by email?

[If yes] Do you have any questions for me?

[If no] Would you like me to walk through the form with you?

[Once done and any questions asked]

I will be using Camtasia, a screen capture software, which will record today session into my local computer. When you are sharing the screen with me, only your voice and your shared
computer screen will be recorded, your face will not be recorded. So please keep the screen sharing on throughout the whole session.

Can you please click “Consent Form” on the top menu bar of the website. The content is the same as the one I sent it to you by email. Please scroll down to the bottom part.

Now, I start recording the study session to capture your verbal consent and the whole study session.

[Once recording is started]

Are you willing to participate in the study? If yes, please say out the statement written in the box of the consent form.

[A After receiving the verbal consent]

Think-aloud Session

Can you please click “Instructions” on the menu bar of the website. Please read the instructions carefully. (link: https://blogs.ubc.ca/ogdresearch/instructions/)

[Wait for the participant to read the instructions]

Have you done reading the instructions? Do you have any questions?

Can you please click the “reply template”. Please read this page, then copy and paste to any word processing software such as Microsoft Word.

[Wait for the participant to copy and paste the reply template]

You can go back to the “Instructions” on the menu bar of the website.

When you are working on the tasks, I would like to say your thoughts out loud that what are you thinking, what are you about to do and why, what works well for you, and what is difficult. I will prompt you or ask you questions to share your thoughts when you are working on the tasks.
You will have 30 minutes to work on the tasks. I will let you know when you have about 10 minutes left and 5 minutes left. Remember, this is not a test, and we are not expecting you to be an expert here – just do your best, and explain as you go.

Are you ready to start?

[The participant to work on the task]

[If the participant is quiet] Tell me more

[If the participant does an action repeatedly] Tell me what you trying to do.

[If the participant uses a certain tool, feature, or function] How helpful is [this tool/feature/function or XXX] in completing your task?

[If the participant seems frustrated/surprised] What just happened?

[If the participant says “I don’t know what to do?”] What would you do if no one was here?

**Questionnaire Session**

Time is up. Please save your working file and email the file to me at the end of the study.

Now, please click “Home” on the menu bar. Please click “Questionnaire” to complete it. You do not need to think-aloud when filling out the questionnaire.

[Wait for the participant to complete the questionnaire]

Thank you for completing the questionnaire.

**Interview Session**

Now, I would like to ask you a few questions about the task that you carried out.

1. First, how did you find the overall experience of using the portal?

2. Can you talk about any challenges or obstacles that you faced when completing the task?
   [If the participant replies “Yes”]

   3a. What did you do to overcome the obstacles? Did you have any strategies to do that?
4a. How helpful or unhelpful were those strategies?

5a. Can you think of any features that could be built into the portal to help you overcome these obstacles?

[If the participant replies “No”]

3b. Could you talk about any strategies you used that helped you to be successful in this task?

4b. Can you think of any features that could be built into the portal that would have been helpful for you, in completing this task, or in using it in general?

5/6. At last, how do you think the open data portals can support people to learn data skills, especially those who may be beginners in working with data?

I stop the recording now.

Honorarium

This is the end of the study. Thank you for participating in this study. I just send you an e-gift card. Please check your email to confirm that you have received the e-gift card.

[Wait for the participant to check email]

Please copy the receipt and attach your files, then email back to me for our records.

Thanks again for participating in this study. I end the session now.
## Appendix B  Codebooks

### B.1  Codebook 1 for coding tools used to find data

<table>
<thead>
<tr>
<th>Code</th>
<th>Tool used</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF1</td>
<td>CoV’s open data portal – Homepage or Catalogue page</td>
</tr>
<tr>
<td>TF2</td>
<td>CoV’s open data portal – Dataset page</td>
</tr>
<tr>
<td>TF3</td>
<td>CoV’s open data portal – VanDashboard page</td>
</tr>
<tr>
<td>TF4</td>
<td>CoV’s open data portal – Chart builder or Map builder page</td>
</tr>
<tr>
<td>TF5</td>
<td>CoV’s open data portal – Help &amp; API page</td>
</tr>
<tr>
<td>TF6</td>
<td>Other websites or tools outside CoV’s open data portal</td>
</tr>
</tbody>
</table>
B.2 Codebook 2 for coding tools used to use data

- **TU1**: CoV’s open data portal – Dataset page – ‘Information’ tool
- **TU2**: CoV’s open data portal – Dataset page – ‘Table’ tool
- **TU3**: CoV’s open data portal – Dataset page – ‘Analyze’ tool
- **TU4**: CoV’s open data portal – Dataset page – ‘Export/API’ tool
- **TU5**: Other websites or tools outside CoV’s open data portal
### Codebook 3 for coding obstacles that participants faced

<table>
<thead>
<tr>
<th>Code</th>
<th>Obstacles faced</th>
<th>Related data literacy competency</th>
</tr>
</thead>
<tbody>
<tr>
<td>F01</td>
<td>Did not know how to search for datasets in the CoV’s open data portal</td>
<td>Find data Search for datasets</td>
</tr>
<tr>
<td>F02</td>
<td>Chose to use WPRG dataset or CBE dataset or another irrelevant dataset instead of ERE-75K dataset</td>
<td>Select the most appropriate dataset</td>
</tr>
<tr>
<td>F03</td>
<td>Could not detect other datasets cannot solve the given problem</td>
<td>Detect whether the data cannot totally or partially solve the given problem</td>
</tr>
<tr>
<td></td>
<td>Could not detect that ERE-75K cannot find the minimum salary of some occupations as employees who earned below $75K per year are not included</td>
<td></td>
</tr>
<tr>
<td>U01</td>
<td>Did not clean data due to multiple representation of data (e.g. librarian II and librarian Ii should be merged for analysis)</td>
<td>Use data Analyze data</td>
</tr>
<tr>
<td></td>
<td>Did not know how to use the selected data analysis tools (table tab, or analyze tab in the CoV’s open data portal, or Microsoft Excel)</td>
<td></td>
</tr>
<tr>
<td>U02</td>
<td>Did not create a table or a graph, or created an improper table or graph (e.g. table without proper headers, graph with inappropriate legends)</td>
<td>Present results for target audience</td>
</tr>
<tr>
<td>U03</td>
<td>Did not cite or improperly cite data sources</td>
<td>Cite data</td>
</tr>
</tbody>
</table>
## Appendix C  Findings of the study

### C.1  Demographics of participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Study level</th>
<th>Faculty</th>
<th>Age Range</th>
<th>Knowledge about OGD</th>
<th>Experience in using OGD website</th>
<th>Experience in using CoV’s open data portal</th>
</tr>
</thead>
<tbody>
<tr>
<td>N01</td>
<td>Master</td>
<td>Applied Science</td>
<td>22-26</td>
<td>Some knowledgeable</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>N02</td>
<td>Undergraduate</td>
<td>Sciences</td>
<td>17-21</td>
<td>Very little knowledgeable</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N03</td>
<td>Undergraduate</td>
<td>Arts</td>
<td>17-21</td>
<td>Little knowledgeable</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N04</td>
<td>Undergraduate</td>
<td>Applied Science</td>
<td>27-31</td>
<td>Some knowledgeable</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>N05</td>
<td>Undergraduate</td>
<td>Land and Food Systems</td>
<td>22-26</td>
<td>Very little knowledgeable</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N06</td>
<td>Master</td>
<td>Applied Science</td>
<td>22-26</td>
<td>Very little knowledgeable</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N07</td>
<td>Master</td>
<td>Education</td>
<td>22-26</td>
<td>No knowledge at all</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N08</td>
<td>Undergraduate</td>
<td>Applied Science</td>
<td>17-21</td>
<td>Very little knowledgeable</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>N09</td>
<td>Undergraduate</td>
<td>Applied Science</td>
<td>22-26</td>
<td>No knowledge at all</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N10</td>
<td>Undergraduate</td>
<td>Sciences</td>
<td>17-21</td>
<td>No knowledge at all</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N11</td>
<td>Undergraduate</td>
<td>Applied Science</td>
<td>17-21</td>
<td>Very little knowledgeable</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N12</td>
<td>Undergraduate</td>
<td>Arts</td>
<td>17-21</td>
<td>No knowledge at all</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N13</td>
<td>Undergraduate</td>
<td>Sciences</td>
<td>17-21</td>
<td>Little knowledgeable</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N14</td>
<td>Master</td>
<td>Sciences</td>
<td>22-26</td>
<td>Little knowledgeable</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
C.2  Self-assessment data literacy survey among each participant

<table>
<thead>
<tr>
<th>Self-assessment of data literacy</th>
<th>Participant number</th>
</tr>
</thead>
<tbody>
<tr>
<td>I can identify different kinds/types of data</td>
<td>N01    N02    N03    N04    N05    N06    N07    N08    N09    N10    N11    N12    N13    N14</td>
</tr>
<tr>
<td>I can describe how data is produced in society</td>
<td></td>
</tr>
<tr>
<td>I can describe a wide range of possible applications of data</td>
<td></td>
</tr>
<tr>
<td>I can find credible data to solve a given problem or need</td>
<td></td>
</tr>
<tr>
<td>I can evaluate the quality of data</td>
<td></td>
</tr>
<tr>
<td>I can identify the limitations of data for a given problem or need</td>
<td></td>
</tr>
<tr>
<td>I can describe the various ways to represent data (e.g. written, numerical or graphic)</td>
<td></td>
</tr>
<tr>
<td>I can interpret data in the form of charts or tables</td>
<td></td>
</tr>
<tr>
<td>I can describe data with metadata</td>
<td></td>
</tr>
<tr>
<td>I can manage data for subsequent reuse</td>
<td></td>
</tr>
<tr>
<td>I can clean and prepare data for analysis</td>
<td></td>
</tr>
<tr>
<td>I can analyze data using tools such as spreadsheets or statistical software</td>
<td></td>
</tr>
<tr>
<td>I can summarize the results of data analysis using various forms of representation (e.g. written, numerical or graphic)</td>
<td></td>
</tr>
<tr>
<td>I can use data ethically (e.g. protect confidentiality, data license)</td>
<td></td>
</tr>
<tr>
<td>I can cite data sources I have used.</td>
<td></td>
</tr>
</tbody>
</table>

Self-assessed high data literacy group (H) or Self-assessed low data literacy group (L)

- Strongly Agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree
### Datasets visited by participants

<table>
<thead>
<tr>
<th>Datasets visited</th>
<th>Participant number</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERE-75K dataset</td>
<td>N01, N02, N03, N04, N05, N07, N08, N09, N10, N11, N12, N13, N14</td>
</tr>
<tr>
<td>It provides the remuneration (i.e. include salary, overtime, gratuity and vacation payouts) and expenses (i.e. include charges such as training, tuition, conferences and travel and professional dues) from employees earning over $75,000 per year</td>
<td></td>
</tr>
<tr>
<td>WPRG dataset</td>
<td>N01, N05, N06, N07, N08, N09, N10, N11, N12</td>
</tr>
<tr>
<td>It provides the hourly rates of pay for all City job classifications and gender breakdown of staff within these classifications, excluding information from the VPL and the Vancouver Police Department</td>
<td>N12</td>
</tr>
<tr>
<td>CBE dataset</td>
<td>N07, N12</td>
</tr>
<tr>
<td>It provides budget and expense of the City Council, including the salary of Mayor and each Council member</td>
<td></td>
</tr>
<tr>
<td>Libraries dataset</td>
<td>N05, N14</td>
</tr>
<tr>
<td>It provides the name, address, latitude, and longitude of each branch of VPL</td>
<td></td>
</tr>
<tr>
<td>Business improvement areas dataset</td>
<td>N06</td>
</tr>
<tr>
<td>It provides the boundary areas of the City’s business improvement area</td>
<td></td>
</tr>
<tr>
<td>Indicator data – Income of Vancouver households</td>
<td>N06</td>
</tr>
<tr>
<td>It provides the income of Vancouver households indicator data in 2010 and 2015</td>
<td></td>
</tr>
<tr>
<td>Film office work areas</td>
<td>N12</td>
</tr>
<tr>
<td>It provides the boundary areas of filming and special event activities within city</td>
<td></td>
</tr>
</tbody>
</table>
C.4 **External websites visited by participants**

<table>
<thead>
<tr>
<th>External websites visited (except study website)</th>
<th>Participant number</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoV’s website</td>
<td>N05, N06</td>
</tr>
<tr>
<td>It provides information and access to CoV government services such as paying parking ticket, finding waste collection schedule, and finding rental housing.</td>
<td></td>
</tr>
<tr>
<td>VPL’s website – Open data research guide</td>
<td>N09, N10, N12</td>
</tr>
<tr>
<td>It provides links to different open data portals such as CoV’s open data portal, DataBC, Government of Canada’s open data portal, and video tutorials on working with open data such as discovering, finding and cleaning open data</td>
<td></td>
</tr>
<tr>
<td>Metro Vancouver’s website – Collective agreements</td>
<td>N11</td>
</tr>
<tr>
<td>It provides the collective agreements between employer and union, such as CoV’s government and CUPE 1004</td>
<td></td>
</tr>
</tbody>
</table>