

**Toward a Model for Human Open Government Data (OGD) Interaction and an
Application for OGD Literacy Taxonomy: A User-centered Perspective**

by

Fanghui Xiao

B.M.S., Capital University of Economics and Business, 2007

M.L.I.S., University of Pittsburgh, 2016

Submitted to the Graduate Faculty of
the School of Computing and Information in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH
SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Fanghui Xiao

It was defended on

November 21st 2022

and approved by

Dr. Daqing He, School of Computing and Information, University of Pittsburgh

Dr. Jacob Biehl, School of Computing and Information, University of Pittsburgh

Dr. Eleanor Mattern, School of Computing and Information, University of Pittsburgh

Dr. David Walker, Western Pennsylvania Regional Data Center, University of Pittsburgh

Center for Social & Urban Research (UCSUR)

Dissertation Director: Dr. Daqing He, School of Computing and Information, University of

Pittsburgh

Copyright © by Fanghui Xiao
2022

Toward a Model for Human Open Government Data (OGD) Interaction and an Application for OGD Literacy Taxonomy: A User-centered Perspective

Fanghui Xiao, PhD

University of Pittsburgh, 2022

Acknowledging the value of transparency and accountability, the development of Open Government Data (OGD) and its portals have been rapidly proliferating around the world. Consequently, massive amounts of government data, from federal to state to local levels, are available via various OGD portals. Also, the emphasis of OGD projects has gradually shifted from a publisher-centered paradigm to a user-centered paradigm, as laws and regulations caught up with policies that make these resources more widely accessible to the public. Thereby, improving data use has become the new major aim of OGD projects. However, extant studies show that users often experience difficulties in finding, understanding, and using government data. Low-level data literacy of individuals was also identified as a major obstacle to using OGD. Within the still-emerging field of human data interaction (HDI), very few studies focus on how users interact with OGD and the fundamental OGD literacy capabilities. Therefore, motivated by the existing challenges of interacting with OGD and the corresponding research gaps in HDI and OGD literacy, this dissertation aims to take a user-centered perspective, relating the relevant models and previous research studies in the two areas, HDI and OGD, to empirically probe into OGD user online interactive behaviors and then to develop a model for human OGD interaction (H-OGD-I). This dissertation also aims to examine contextualized user challenges of interacting with OGD, pinpoint user literacy challenges and platform design barriers to identify the fundamental OGD literacy capabilities that enable users to use OGD based on the proposed H-OGD-I model, and finally develop a taxonomy for OGD literacy capabilities.

This dissertation focuses on the users of local-level rather than federal- or state-level OGD portals. Previous studies claimed that the OGD is mainly collected at local-level, thus, supporting local-level OGD is to support the success of OGD as a whole. Also, a local-level OGD portal is critical because of its closer connection with local organizations,

neighborhoods, and communities, which more directly impacts a citizen's daily life and neighborhood. Accordingly, the users from three local-level OGD portals were studied in this dissertation: the city of Philadelphia (OpenDataPhilly), the Western Pennsylvania Regional Data Center (WPRDC), and the city of Boston (Analyze Boston). The targeted users are the non-expert end users who have experience interacting with OGD. An end user refers to an individual who uses the OGD directly rather than consuming the effects of the OGD application, e.g., by using transportation apps. To achieve the objectives and answer the research questions of this dissertation, five sub-studies with mixed research method design were conducted, including Study 1: observing users' OGD accessing behavioral patterns, Study 2: identifying users' individual OGD behaviors, Study 3: exploring users' cascading OGD behaviors, Study 4: examining user challenges in each H-OGD-I stage, and Study 5: investigate fundamental OGD literacy capabilities.

This dissertation first contributes to the HDI field by providing a model for H-OGD-I with a deep insight into user OGD behaviors. This new H-OGD-I model accounts for the stages and user behaviors when interacting with OGD, and therefore it advances the understanding of the complexity of OGD user behaviors. Also, the newly discovered behaviors in the stages of Sensemaking and Sharing make this H-OGD-I model more comprehensive. In addition, this model is expected to be generalized to structured research data due to the similarity of the infrastructure of structured data, which may inspire researchers in the field of research data management to understand their users' behaviors. Furthermore, this H-OGD-I model can assist OGD interface designers in defining more effective user interactions that help users find, acquire, and make sense of data, as well as facilitate librarians and instructors to develop OGD literacy capabilities. Additionally, the second main contribution – the taxonomy of OGD literacy capabilities was developed based on the H-OGD-I model and two empirical studies, which explored and identified the fundamental capabilities that are needed in each stage of OGD interaction. This taxonomy can contribute to guiding the design of education programs, namely, data literacy-related curricula or workshops, to enhance users' OGD literacy. Given few research studies exist focusing on OGD literacy, this OGD literacy taxonomy can contribute to filling this research gap in data literacy.

Overall, this dissertation provides a holistic view and a deep insight into human OGD

interaction (H-OGD-I model) and offers a corresponding application of OGD literacy capabilities (taxonomy). The comprehensive H-OGD-I model makes a theoretical contribution to the field of HDI and the subsequent taxonomy of OGD literacy capabilities makes a practical contribution, which advocates for improving people's data literacy skills.

Table of Contents

1.0 Introduction	1
1.1 Background of the Study	1
1.2 Key Concepts	4
1.2.1 Open Government Data (OGD)	4
1.2.2 Open Government Data Portal	6
1.2.3 Human Data Interaction	6
1.2.4 Open Government Data Literacy	7
1.3 Problem Statement	7
1.4 Research Objectives	10
1.5 Research Motivation	11
2.0 Literature Review	13
2.1 Identified Challenges and Ongoing Efforts of OGD Project	14
2.1.1 The Challenges and Ongoing Efforts of Finding OGD	14
2.1.1.1 The Challenges of Finding OGD.	14
2.1.1.2 The Scholarly Work on Improving Findability of OGD.	16
2.1.2 The Challenges and Ongoing Efforts of Understanding OGD	18
2.1.2.1 The Challenges of Understanding OGD.	18
2.1.2.2 The Scholarly Work on Improving Understandability of OGD.	19
2.1.3 The Challenges and Ongoing Efforts of Using OGD	20
2.1.3.1 The Challenges of Using OGD.	20
2.1.3.2 The Scholarly Work on Improving Usability of OGD.	22
2.2 Human Information Interaction	24
2.2.1 The Concepts of Human Information Interaction	25
2.2.2 Two HIB models: Big Six and Information Search Process	26
2.2.3 Information Needs and Search Strategy	28
2.3 Theories and Models on Sensemaking	30

2.3.1	Dervin’s Sensemaking Theory	30
2.3.2	Pirrolli & Card’s Conceptual Model of sensemaking	32
2.4	Human Data Interaction & HDI in Open Government Data	35
2.4.1	The Framework for Human Structured-data Interaction	35
2.4.2	Process Framework for Users’ Activities for Using OGD	37
2.4.3	HDI in Open Government Data	38
2.5	The Development of Data Literacy	39
2.5.1	The Beginning Phase (An Information Literacy and Statistical Literacy Phase)	40
2.5.2	The Second Phase (A Technical and Skill-based Phase)	41
2.5.3	The Third Phase (A Conceptualization Phase)	42
3.0	Research Questions, Research Design and Conceptual Model	45
3.1	Research Questions	45
3.2	Overview of Research Design	46
3.3	Conceptual Model for H-OGD-I	48
4.0	Dissertation Studies	52
4.1	Research Sites and Research Population	52
4.2	Data Collection for the Three Datasets	53
4.2.1	Transaction Log Data Collection	54
4.2.2	Online Forum Posts Data Collection	55
4.2.3	Interview Data Collection	57
4.3	Study 1: Observing Users OGD Accessing Behavior Patterns	58
4.3.1	Data Processing and Analysis	58
4.4	Study 2: Identify Users’ Individual OGD Behaviors	60
4.4.1	Data Analysis	60
4.4.2	Coding Schema	61
4.5	Study 3: Explore Users’ Cascading OGD interacting behaviors	64
4.5.1	Data Analysis	64
4.6	Study 4: Examine User Challenges in each H-OGD-I stage	65
4.6.1	Data Analysis	65

4.6.2 Coding Schema	66
4.7 Study 5: Investigate Fundamental OGD Literacy Capabilities	67
4.7.1 Data Analysis	67
4.7.2 Coding Schema	68
5.0 Results	71
5.1 What are the Typical User Behaviors When Interacting with OGD? (RQ1) .	71
5.1.1 What Types of Tasks are Performed When Users Interact with OGD? (RQ1.1)	71
5.1.2 What are the Users' Accessing Behavioral Patterns When Interacting with OGD and Its Portal Websites? (RQ1.2)	73
5.1.2.1 Commonly Used Channels for Accessing OGD Portals.	73
5.1.2.2 User Preferences on Browsing and Searching.	79
5.1.2.3 User Preferences on Data Entries.	80
5.1.3 What are the Typical User Behaviors in Each Interacting Stage? (RQ1.3)	82
5.1.3.1 Task Preparation Stage.	85
5.1.3.2 Data Forage Stage.	86
5.1.3.3 Data Sensemaking Stage.	89
5.1.3.4 Data Use Stage.	91
5.1.3.5 Data Share Stage.	93
5.1.4 The Evaluation of the Behavior Model for Human OGD Interaction .	94
5.1.5 Outcome1: The Model for Human OGD Interaction	95
5.2 What are the Contextualized Challenges Users Face in Each H-OGD-I stage? (RQ2)	97
5.2.1 The Challenges in the Context of the <i>Task Preparation Stage</i>	98
5.2.1.1 The Challenges in the Context of the <i>Framing Data-centric Question Process</i>	98
5.2.2 The Challenges in the Context of the <i>Data Foraging Stage</i>	99
5.2.2.1 The Challenges in the Context of the <i>Finding Process</i>	99
5.2.2.2 The Challenges in the Context of the <i>Acquiring Process</i>	102
5.2.2.3 The Challenges in the Context of the <i>Scrutinizing Process</i>	103

5.2.3	The Challenges in the Context of <i>Data Sensemaking Stage</i>	105
5.2.3.1	The Challenges in the Context of the <i>Understanding Process</i> . .	105
5.2.3.2	The Challenges in the Context of the <i>Schematizing Process</i> . .	107
5.2.4	The Challenges in the Context of the <i>Data Use Stage</i>	108
5.2.4.1	The Challenges in the Context of the <i>Preparing Process</i>	108
5.2.4.2	The Challenges in the Context of the <i>Analyzing Process, Representing Results Process, and Communicating Process</i>	109
5.2.5	A Summery for RQ2	111
5.3	What are the Fundamental OGD Literacy Capabilities that Enable Users to Use OGD? (RQ3)	112
5.3.1	The Fundamental Action-specific OGD Literacy Capabilities in the <i>Task Preparation Stage</i>	114
5.3.1.1	The Capabilities in the Process of <i>Exploring Existing Datasets</i> . .	114
5.3.1.2	The Capabilities in the Process of <i>Framing Data-centric Questions</i>	114
5.3.1.3	The Capabilities in the Process of <i>Developing Data Needs</i> . . .	115
5.3.2	The Fundamental Action-specific OGD Literacy Capabilities in the <i>Data Foraging Stage</i>	117
5.3.2.1	The Capabilities in the Process of <i>Finding Data</i>	118
5.3.2.2	The Capabilities in the Process of <i>Scrutinizing Data</i>	119
5.3.2.3	The Capabilities in the Process of <i>Acquiring Data</i>	121
5.3.3	The Fundamental Action-specific OGD Literacy Capabilities in the <i>Data Sensemaking Stage</i>	123
5.3.3.1	The Capabilities in the Process of <i>Understanding Data</i>	123
5.3.3.2	The Capabilities in the Process of <i>Schematizing Data</i>	125
5.3.4	The Fundamental Action-specific OGD Literacy Capabilities in the <i>Data Use Stage</i>	128
5.3.4.1	The Capabilities in the Process of <i>Preparing Data</i>	129
5.3.4.2	The Capabilities in the Process of <i>Analyzing data, Representing Results, Interpreting, and Communicating Data</i>	130

5.3.5	The Fundamental Action-specific OGD Literacy Capabilities in the <i>Data Sharing</i> Stage	134
5.3.6	The Fundamental Action-holistic OGD Literacy Competencies	135
5.3.6.1	Conceptual Foundations.	135
5.3.6.2	Advanced Actions.	136
5.3.6.3	Dispositions.	138
5.3.7	The Evaluation of the Taxonomy for OGD Literacy Capabilities . . .	140
5.3.8	Mapping to Existing Data Literacy Capabilities	141
5.3.9	Outcome 2: The Taxonomy for OGD Literacy Capabilities	145
6.0	Discussion and Implications	150
6.1	Discussion of Results	150
6.1.1	OGD Users' Accessing Behavior Patterns	150
6.1.2	OGD Users' Challenges When Interacting with Data	152
6.1.3	The Fundamental Capabilities for OGD Literacy	153
6.2	Discussion of Research Design	155
6.3	Implications of H-OGD-I Model	157
6.3.1	Design Implications	157
6.3.2	Theoretical Implications	158
6.4	Implications of the Taxonomy of OGD Literacy Capabilities	159
6.4.1	Design Implications	159
6.4.2	Theoretical Implications	160
6.5	Design Implications of the Forum Data Analysis	161
7.0	Conclusion	162
7.1	Significance and Contribution	162
7.2	Limitations	164
7.3	Future Work	165
7.3.1	Improving OGD Use	166
7.3.2	Promoting Data Equality.	167
Appendix A. Interview Participant Recruitment Email		169
Appendix B. Interview Protocol		171

Appendix C. The Model for Human OGD Interaction (H-OGD-I)	174
Appendix D. The Taxonomy for OGD Literacy Capabilities	178
References	182

List of Tables

1	Browsing by Different Entries	59
2	The Coding Schema for OGD User Challenges	66
3	The Coding Schema for OGD Literacy Capabilities	68
4	OGD Literacy Capabilities in Holistic Process	70
5	Tasks and Task Types	73
6	Typical User Behaviors Identified by the Analysis of Forum Posts and Interview Data	83
7	Four Aspects that OGD Users Mainly Scrutinized	88
8	OGD Literacy Capabilities in the Stage of <i>Task Preparation</i>	117
9	OGD Literacy Capabilities in the Stage of <i>Data Foraging</i>	122
10	OGD Literacy Capabilities in the Stage of <i>Data Sensemaking</i>	128
11	OGD Literacy Capabilities in the Stage of <i>Data Use</i>	133
12	OGD Literacy Capabilities in the Stage of <i>Data Sharing</i>	134
13	Competencies-Conceptual Foundations	136
14	Competencies-Advanced Actions	138
15	Competencies-Dispositions	140
16	OGD Literacy Capabilities Mapping	142
17	The Fundamental Action-specific OGD Literacy Capabilities	178
18	The Fundamental Action-holistic OGD Literacy Competencies	181

List of Figures

1	Preliminary ecosystem of OGD programs (Dawes et al., 2016)	2
2	Foundations of Open Government Data (Gonzalez-Zapata & Heeks, 2015)	5
3	Pew Research Center: Survey	8
4	The Core Concepts of HII	25
5	Kuhlthau’s Model of the Information Search Process (ISP) (Kuhlthau, 1991) . .	28
6	Dervin’s Sensemaking Theory (Dervin, 2005)	31
7	Pirrolli & Card’s Conceptual Model of Sensemaking in Intelligence Analysis (Pirrolli & Card, 2005)	33
8	Framework for Interacting with Structured Data. (L. Koesten et al., 2017)	35
9	The User Process with Activities and Variations. (J. Crusoe & Ahlin, 2019) . . .	37
10	Overview of Research Design	46
11	The Proposed Conceptual Model of HDI	49
12	The Overview of Data Collection	54
13	Homepage of Discussion Group for Open Data and Government Transparency in Philadelphia	55
14	An Example of Posts	56
15	An Example of Group Data Collection	56
16	The Distribution of Channels for Accessing the Portals	74
17	The Distribution of Channels for Accessing the Portals—OpenDataPhilly	75
18	The Distribution of Channels for Accessing the Portals—WPRDC	76
19	The Distribution of Channels for Accessing the Portals—Analyze Boston	76
20	The Total Conversion Rates of Portals	78
21	The Conversion Rates of Each Channel Among Portals	79
22	The Accessing Behaviors among the Three OGD Portals	80
23	The Data Entries in the Three OGD Portals	81
24	Behaviors Manifested by Each Participant	84

25	H-OGD-I Model — Task Preparation Stage	85
26	H-OGD-I Model — Data Forage Stage	87
27	H-OGD-I Model — Data Sensemaking Stage	90
28	H-OGD-I Model — Data Use Stage	91
29	Finalized Model for H-OGD-I	96
30	The Contextualized Interactive Challenges User Encountered	98
31	The Sub-categories of <i>Find</i> Challenges	100
32	The Sub-categories of <i>Scrutinize</i> Challenges	104
33	The Sub-categories of <i>Understand</i> Challenges	106
34	The Sub-categories of <i>Prepare</i> Challenges	109
35	The Distribution of the OGD Literacy Capability in Each Behavioral Process . .	113
36	The Fundamental Action-holistic OGD Literacy Competencies	146
37	The Taxonomy for OGD Literacy Capabilities - Task Preparation	146
38	The Taxonomy for OGD Literacy Capabilities-Data Foraging	147
39	The Taxonomy for OGD Literacy Capabilities-Data Sensemaking	147
40	The Taxonomy for OGD Literacy Capabilities - Data Use	148
41	The Taxonomy for OGD Literacy Capabilities - Data Sharing	149
42	The Model for Human OGD Interaction (H-OGD-I)	174

1.0 Introduction

1.1 Background of the Study

Recent dramatic increases in the ability to generate, collect, and use datasets have inspired numerous academic and policy discussions regarding the emerging field of human data interaction (HDI). This dissertation aims to explore HDI in the open government data (OGD) domain and investigate ways HDI can further contribute to OGD promotion. The Organization for Economic Cooperation and Development (OECD) conceives OGD as “a philosophy- and increasingly a set of policies - that promotes transparency, accountability, and value creation by making government data available to all.”¹ This notion of OGD suggests that with publishing OGD, public bodies become more transparent and accountable to citizens; the free OGD use can promote social engagement, citizen-centric services, and business creation and innovation if the OGD can be efficiently used. After more than one decade, the development of OGD and its portals have been multiplying rapidly worldwide; for instance, in Europe, the Public Sector Information (PSI) Directive in 2003; in the U.S, the Open Data initiatives in 2009 and the OPEN Government Data Act in 2019; and the International G8 Open Data Chapter in 2013.

OGD involves multiple stakeholders, who may seek different outcomes. An OGD ecosystem model proposed by Dawes et al. (2016) demonstrates the operational process of implementing the OGD program and implicates the composition of OGD stakeholders (see Figure 1). This model describes three primary stakeholder groups: (1)government leaders and organizations responsible for OGD programs, comprising elected officials, data administrators (Dawes et al., 2016), legislators, transparency and ethics commission, public sector practitioners, international organizations (Gonzalez-Zapata & Heeks, 2015); (2)direct OGD users who comprise transparency advocates, including expert data analysts and members of the civic technology community (Dawes et al., 2016), civil society activists, academics, journalist(Gonzalez-Zapata & Heeks, 2015); and (3)the beneficiaries of OGD use, which refers to

¹open government data, <https://www.oecd.org/gov/digital-government/open-government-data.htm>

both individuals and organizations who adopt, buy, and use the products and services that OGD has made possible. While Dawes et al. (2016) and Gonzalez-Zapata & Heeks (2015) argue that individual citizens are not frequent users of OGD, Ubaldi (2013) stated, citizens are the stakeholders of OGD precisely because public participation and social engagement are central missions of OGD. Additionally, the stakeholder groups can overlap as they engage in different tasks. For example, when a public sector practitioner uses OGD to formulate a work plan for the next year, she becomes a direct user.

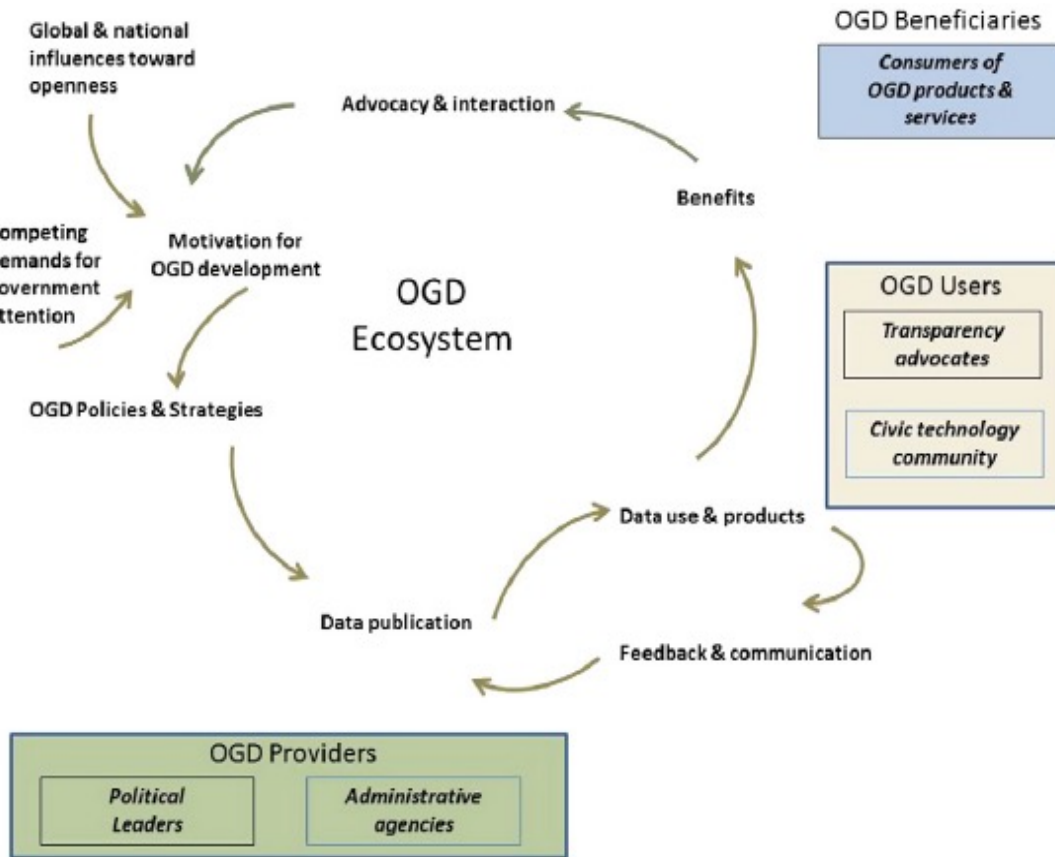


Figure 1: Preliminary ecosystem of OGD programs (Dawes et al., 2016)

There is ample research that discussed the values of OGD. In government, the focus is on providing better-customized services by using OGD and improving transparency and accountability (Assaf et al., 2015; Gonzalez-Zapata & Heeks, 2015; Ubaldi, 2013). OGD could also spur innovations that improves citizens' life. For example, in 2008, Washington DC held

an innovation contest that had citizens use the city’s public datasets, such as real-time crime feeds, school test scores, and poverty indicators, to create software applications to engage civic literacy and promote a better understanding of the local political system. With an operating cost of \$0.05 million, the “Apps for Democracy” contest yielded 47 apps with a market value of \$2.3 million. This idea of using OGD to create Apps for Democracy has inspired similar movements in over 50 countries and cities around the world.² OGD may help individuals make better decisions as well (Ubaldi, 2013). For instance, prospective homebuyers could use OGD to find more information about neighborhoods and schools. Besides these political and social values, OGD is essential to economy (Wirtz et al., 2022). Manyika et al. (2013) asserted that, when fully used, open data has the potential to generate USD \$3 million globally and annually. Another report by Gruen et al. (2014) illustrated the potential economic value of reinvigorating the open data agenda in Australia and the G20. This could increase G20 output by around USD \$13 trillion over the next five years, including the themes of trade, finance, anti-corruption, employment, energy, and so forth (Gruen et al., 2014). To reach the full potential of OGD, including its political, social, and economic values, the data should be maximally used by the public. That is why the ultimate mission of local OGD portals is to improve data use (Xiao, Lyon, et al., 2018).

The increase in generating, collecting, and using datasets has inspired numerous discussions regarding how humans interact with data, with different definitions of the term “human data interaction” (HDI) itself. HDI is an emerging area of study (Crabtree & Mortier, 2015), and it deserves to be treated as its own field (Haddadi et al., 2013). In this dissertation, I interpret the notion of HDI from the lens of information science (IS). Within the context of OGD, the concept of HDI is defined as an area of research that investigates how humans interact with both raw data and structured data, focusing on the relationships between humans and data, including users’ tasks that motivate their data use, and the behaviors of how users look for, evaluate, scrutinize, acquire, make sense of, and use data. This understanding of HDI within the context of OGD has the potential to be applied to other structured data in different contexts, such as research data.

In addition, a thorough review of the literature on the challenges of interacting with OGD

²Apps For Democracy: An Innovation Contest, <https://isl.co/work/apps-for-democracy-contest/>

revealed that lacking the necessary knowledge or skills is a crucial factor that inhibits users from accessing (Liu & Jagadish, 2009) and using OGD (C. Martin, 2014). These difficulties require users’ data literacy skills to be addressed. Existing studies also detected the rise of new issues, such as data inequality. There is “a growing gap between those who can work effectively with data and those who cannot” (D’Ignazio, 2017). The inequality can take two forms: 1) the inequality between those who are proficient in data-domain knowledge, e.g., storage and collection, and those who are not (Andrejevic, 2014) and 2) the inequality between those who are proficient in the technical skills required to effectively work with data and those who are not (D’Ignazio, 2017). To combat these inequalities, it is essential to significantly improve citizens’ data literacy. Although data literacy has been discussed for decades (Koltay, 2015; Prado & Marzal, 2013), few studies have specifically investigated data literacy related to OGD (in short, OGD literacy). This dissertation aims to explore and identify the needed fundamental OGD literacy capabilities for non-expert users from a user-centered lens.

1.2 Key Concepts

In this section, I discuss four core concepts that form the basis of this dissertation: 1) open government data, 2) open government data portals, 3) human data interaction, and 4) open government data literacy. These concepts are significant to understand this dissertation.

1.2.1 Open Government Data (OGD)

The concept of OGD can be complex, involving three essential factors - open government, government data, and open data (Figure 2) (Gonzalez-Zapata & Heeks, 2015). Open government, initially, refers to “the notion that people have the right to access the documents and proceedings of government” (Lathrop & Ruma, 2010). With the development of open government movement, the understanding of open government evolved, which signifies that citizens can not only access the documents and proceedings, but also can par-

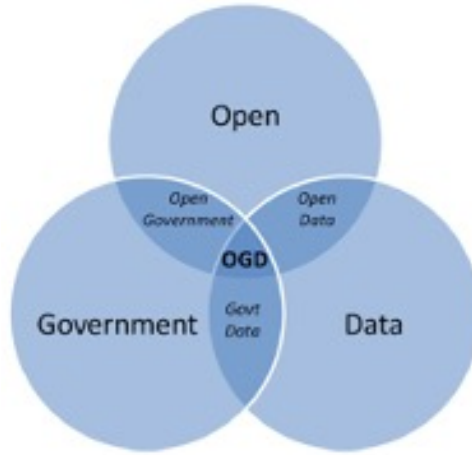


Figure 2: Foundations of Open Government Data (Gonzalez-Zapata & Heeks, 2015)

ticipate in some operations (Lathrop & Ruma, 2010). Open government is expected to produce greater governmental efficiency, transparency, and accountability,³ and OGD is a part of the open government movement (Lathrop & Ruma, 2010). Government data denotes "any information, document, media, or machine-readable material regardless of physical form or characteristics, that is created or obtained by the government in the course of official government business." (Code_of_Federal_Regulations, 2015). Currently, in addition to the government data, a number of data produced or obtained by public bodies and non-profit organizations (not strictly government data) have also been available on OGD portals. Open data refers to the data "must be published in an accessible format, with a licence that permits anyone to access, use and share it."⁴ Another related key concept is open, that indicates that "anyone can freely access, use, modify, and share for any purpose."⁵

Open Government Data. Based on the four essential definitions discussed above, this dissertation defines open government data as any structured raw data generated and published by governments, across all government levels, public bodies, as well as non-profit organizations, and are made available to the public to freely access, reuse, and redistribute without copyright restrictions. Also, the data must be published in both machine-readable

³open government, <https://www.oecd.org/gov/open-government/>

⁴Open data institute, <https://theodi.org/article/open-data-means-business/>

⁵Open Knowledge Foundation, <http://opendefinition.org/>

and human-accessible formats.

Given that most OGD on OGD portals are structured, this proposed definition focuses on the type of structured data, which may cause a limitation when the OGD is unstructured in the future. *Structured data* refers to the data “that can be easily organized, stored and transferred in a defined data model, such as numbers/text set out in a table or relational database that have a consistent format” (Kitchin, 2014).

1.2.2 Open Government Data Portal

An OGD portal can be defined as “an official web-portal launched at the federal or local level aimed at making certain types of governmental datasets publicly accessible via internet in a machine-readable format” (Kassen, 2013). Specifically, an OGD portal supports search functionalities to facilitate finding datasets of interest, and contains metadata records of datasets published for re-use. Application Programming Interfaces (APIs) are also often available, offering direct and automated access to data for software applications.⁶ Nowadays, OGD portals extend the various functions to promote legibility and comprehension, which aligns with the primary aim of OGD portals that is to make OGD easy to be found, accessed, understood, and used.

1.2.3 Human Data Interaction

Based on existing models of human information behaviors, L. Koesten et al. (2017) developed a framework for human structured-data interaction, comprising task, search, evaluate, explore and use. Also, the concept of human information interaction (HII) from Fidel (2012)’s research is defined as the interaction between people and information with its multiple forms and purposes, focusing on the relationships between people and information. In this dissertation, by referencing these two concepts, I propose the concept of human data interaction within the context of OGD is defined as an area of research that investigates how humans interact with both raw data and structured data, focusing on the relationships between humans and data, including users’ tasks that motivate their data use, and the behaviors of how

⁶Open Data Portal, <https://ec.europa.eu/digital-single-market/en/open-data-portals>

users look for, evaluate, scrutinize, acquire, make sense of, and use data. This understanding of HDI in the OGD domain guides my dissertation studies.

1.2.4 Open Government Data Literacy

Gray et al. (2018) proposed the idea of “global data literacy” and suggested considering data infrastructure literacy as a site for ongoing public engagement and experimentation. OGD is one of the ways in which data can promote transparency and enhance public engagement. By referencing this idea and the literature review associated with data literacy, in this dissertation, OGD literacy is defined as the *abilities* to integrate social or individual context to ask the right questions, locate needed data, make sense of data, and use data to make a data-driven decision or to communicate, such as telling a data story.

1.3 Problem Statement

The arguments of opening government information have been in existence for decades (Cranefield et al., 2014; Magalhaes & Roseira, 2017; Masip-Bruin et al., 2013). Scholars, researchers, and industry practitioners still argue that governments should release their data and make the data free of use to everyone for the public good. Such a move would promote government transparency and accountability. Additionally, encouraging the use and reuse of the OGD will generate increased social and economic value (Wirtz et al., 2022). For example, since 2020, the world has been experiencing the COVID-19 pandemic, and it has continued for more than three years. This tremendous and unforeseen event led to an increased demand for corresponding OGD. By accessing COVID-19 data made available on OGD platforms, public members have created tracking maps to visualize the data for public use; interested citizens have started tracking infection and hospitalization rates to understand the current situation. However, users still face challenges that block them from a richer use of available data. User challenges in turn complicate efforts by government and public bodies to increase transparency and accountability, two of the main goals of OGD.

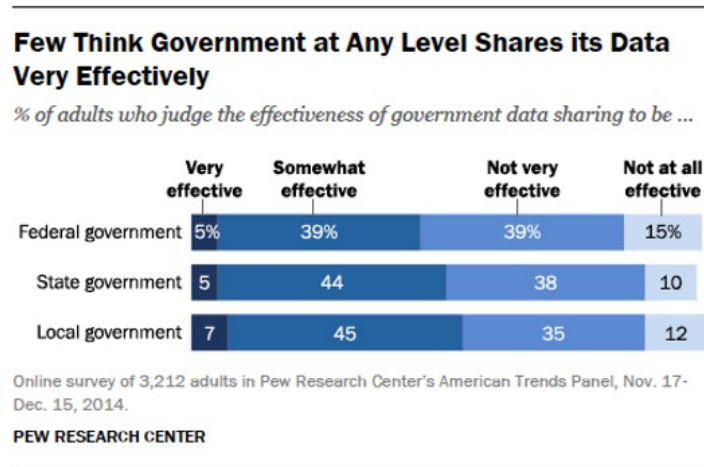


Figure 3: Pew Research Center: Survey

With the explosion in the number of available OGD data portals around the world, researchers have begun to examine the performance of the OGD project, including whether or not the results satisfy the expectations of the advocates, whether everyone is able to use these data efficiently, and whether or not the claimed values are fully achieved. Unfortunately, the answer to these concerns is in the negative. A 2014 survey that examines Americans' views about OGD was conducted by Pew Research Center. Over 3,000 Americans participated in this survey, and the discussed OGD covered the federal, state, and local levels. This survey found that even though the U.S. public is optimistic about OGD, very few think “agencies are doing a great job of providing useful data.” Shown as in figure 3, few people think that governments are “very effective” at sharing data with the public, and 10-15% of respondents even think that governments are “not at all effective” at sharing data.

Although this survey did not discuss why respondents think governments are not effective at sharing data, the reasons could be identified by several seminal studies. These studies identified the challenges of using OGD that exist through the entire OGD interaction process, from finding to understanding and to using the data. Users claimed that “data is hard to find” (Lammerhirt, 2017; Xiao et al., 2019; Zuiderwijk, Janssen, et al., 2012). They described that multiple queries are required to access the needed dataset due to bad naming or website indexing (Li et al., 2022; Zuiderwijk, Janssen, et al., 2012), or that finding datasets in

many practical situations was not always straightforward and may need to be collected from different sources (L. Koesten et al., 2017; Lammerhirt, 2017).

Researchers also found that compared to finding OGD, understanding OGD is more difficult for users (J. Crusoe et al., 2019). Users complained that the platforms lack context information to interpret data and lack explanations for the meaning of data, and users may lack the knowledge to make sense of data (Janssen et al., 2012; Mutambik et al., 2021; Osagie et al., 2017). If users cannot find or make sense of data, they cannot use the data, let alone create value out of these data.

Furthermore, even if users successfully find the data that they need, they could still encounter barriers using the data, including poor data quality, the lack of OGD use standards (Ruijter & Meijer, 2020), and complex OGD platforms (Osagie et al., 2017). The lack of accuracy and timeliness of data is a big issue of data quality (Janssen et al., 2012; Lee & Kwak, 2012; Vetrò et al., 2016). For instance, if there are missing values in the dataset, the user may be forced to give up on using it. In addition, some OGD information is available only in PDF form, a non-machine-readable format, and therefore unable to be reused (Janssen et al., 2012). Another barrier of using OGD is a lack of adoption of OGD use standards, including the variables used, the standards of metadata and classifications, as well as concepts, which could significantly affect the extent to which data can be moved around, linked and made use of and shared (Commission, 2017; Huijboom & Van den Broek, 2011; Nikiforova & McBride, 2021)

Besides the challenges caused by platform design related to interface and content infrastructure, interacting with OGD also has a unique user difficulty. In order for a user to effectively use the data, the user must have a certain level of OGD literacy (Andrejevic, 2014; D'Ignazio, 2017; Lněnička et al., 2022). A thorough review of the literature on the challenges of interacting with OGD confirms that lacking the necessary knowledge or skills is a crucial factor that inhibits users from accessing and using OGD (Liu & Jagadish, 2009; C. Martin, 2014; Xiao et al., 2022). An essential factor that distinguishes data literacy from information literacy is that information literacy focuses wholly on reading and comprehension skills. However, in addition to reading and comprehension skills, data literacy also requires a certain level of computational skills (statistical and technical). This data

characteristic can extremely impede non-expert users from using data. Extant studies point out that most potential OGD users lack OGD literacy skills (Boyчук et al., 2016; Li et al., 2022), especially computational skills (Xiao et al., 2022). Even though OGD portals can offer interactive tools to facilitate users to manipulate data, they cannot cater to all users' needs when a user is without OGD literacy.

All these obstacles can cause huge losses not only for the public but also for governments. With the breaking down of the barriers to OGD, the public can be empowered to make better data-informed decisions and to spark innovations that are expected to make citizens' life easier. Further, with better use of OGD, governments could win the trust of citizens, and therefore, help facilitate administrative governance. Even though existing studies have identified user challenges in interacting with OGD, research studies are still lacking the support to solve these challenges, especially from the angle of investigating OGD users. Therefore, this dissertation work intends to take a stand from a user-centered lens, to explore user OGD interactive behaviors, pinpoint and understand the contextualized challenges that the users face, and then investigate the fundamental OGD literacy capabilities, to ultimately promote OGD use.

1.4 Research Objectives

The overarching goal of my dissertation is to promote OGD use through understanding OGD users, including user data behaviors and the corresponding connections with their OGD literacy. Accordingly, the research goal is to uncover and understand users' online OGD behaviors and the contextualized user challenges when interacting with data and ultimately contribute a model for human OGD interaction (H-OGD-I) to the field of HDI. Additionally, based on the proposed H-OGD-I model, this study aims to develop a taxonomy for OGD literacy capabilities.

To proceed, four objectives are identified:

Objective 1: to explore and understand OGD users' online data interacting behaviors, including 1) explore what types of tasks are performed when users interact with OGD, 2)

observe accessing behavior patterns, 3) identify users' individual behaviors, and 4) investigate users' cascading interacting behaviors.

Objective 2: to develop a model for human OGD interaction (H-OGD-I) based on the understanding of Objective 1.

Objective 3: to examine and contextualize user challenges in each H-OGD-I stage, and pinpoint user literacy challenges and platform design barriers.

Objective 4: to explore the fundamental OGD literacy capabilities that enable users to use OGD based on the proposed H-OGD-I model and develop a taxonomy for OGD literacy capabilities.

1.5 Research Motivation

As identified in the problem statement, the existing obstacles greatly hinder potential OGD users from using the data to achieve its full potential. Therefore, my first and foremost motivation for this dissertation is to promote OGD use to generate greater value in various fields. With the rapid development of OGD, various challenges have arisen along with it, which conceivably restrain the widespread and effective OGD use. Given this, I aim to address this issue from a user-centered perspective, understanding users' OGD interactive behaviors, contextualizing users' challenges within each stage when interacting with OGD, and then developing a corresponding taxonomy of OGD literacy capabilities. Proposing this OGD literacy taxonomy is motivated by the goal of promoting data equality. Recent studies address the issue of data inequality that has become more prominent with the explosive growth of OGD. There is "a growing gap between those who can work effectively with data and those who cannot" (D'Ignazio, 2017). This taxonomy can assist in designing educational programs and materials to enhance users' OGD literacy, which is key to promoting data use and equality.

In addition, I aim to fill the gap in the literature on human data interaction (HDI). People generate data every day consciously or subconsciously and use data in various manners, such as direct use or by using data conglomerates, e.g., a weather app. Data is currently used in

multiple scenarios and areas, such as forming policy plans, conducting scientific studies, and guiding personal decisions. Therefore, scrutinizing and understanding how people interact with data—their behaviors, challenges, and concerns—will be extremely helpful in designing systems to best facilitate data use. However, few empirical studies have been conducted to explore pertinent research inquiries. In fact, the area of HDI is still an emerging field, and many scholars believe it is on the way to becoming its own distinct field of study (Crabtree & Mortier, 2015; Mortier et al., 2013; Victorelli et al., 2020). With these motivations, I aim to develop a comprehensive model of human OGD interaction to contribute to the HDI field and serve as a theoretical foundation and guidance for future research.

Last but not least, I have always been interested in human information/data interaction. Also, when I learned about the meaning and values of “open data,” I was captivated by the idea of relating the two fields to study how humans interact with open data, and OGD in particular. The public sectors make a great effort to make data available to the public and strive toward the engagement of each individual, which provides an extensive research space to investigate how individuals interact with OGD. Therefore, my research attempts to understand user experiences associated with OGD from a user-centered lens.

2.0 Literature Review

This literature review consists of five sections: 1) the challenges concerning finding, understanding, and using OGD, and the corresponding scholarly work about ongoing efforts in offering solutions to these challenges, 2) the concepts and models in the field of human information interaction (HII), 3) the theories or models in the field of sensemaking, 4) the models in the fields of HDI and the empirical studies applying HDI models to work in the OGD field, and 5) the development of data literacy.

Researchers with a pragmatic worldview pay attention to the research problem and apply pluralistic approaches available to understand and address the problem (Creswell, 2014). In this view, two primary goals of this dissertation were set up, exploring and understanding users' OGD interactive behaviors, and identifying the fundamental data literacy capabilities for OGD users. I first went over the literature on OGD users' challenges in interacting with the data. Therefore, section 2.1 is dedicated to examining the challenges concerning finding, understanding, and using OGD and the corresponding scholarly work about ongoing efforts in offering solutions to these impediments. This literature exploration crystallized the research problems and provided a fundamental background for this dissertation.

Next, this dissertation explored the concepts of HDI to address the research problem. HDI is inherently a multidisciplinary area intertwined with other research fields, including but not limited to HCI, data visualization, and Computer-Supported Cooperative Work (CSCW). In this dissertation, I consider HDI from the lens of information science and combine it with a sensemaking model to propose an initial conceptual model of HDI. Given this, in section 2.2 and 2.3, I reviewed the representative theories or models in the areas of HII/HDI and sensemaking, which build the groundwork for the revised conceptual model devised in this dissertation. After going over the theoretical literature related to HDI, HII, and sensemaking, the research studies on the applications of HDI models in OGD practices were examined (section 2.4). At last, with the second goal of this dissertation, to explore and identify the fundamental data literacy capabilities for OGD users, I reviewed the research studies discussing the development of data literacy, including the concept of data literacy

and the proposed competencies, displayed in section 2.5.

This Chapter depicts a complete picture of the current research status of the relevant themes brought up in this dissertation study, will all shaped the direction of this project.

2.1 Identified Challenges and Ongoing Efforts of OGD Project

2.1.1 The Challenges and Ongoing Efforts of Finding OGD

2.1.1.1 The Challenges of Finding OGD. The challenge of finding OGD has been identified by many researchers for a decade. Many researchers claimed that poor findability of OGD is a decisive impediment to OGD adoption (Swamiraj & Freund, 2015; Zuiderwijk, Janssen, et al., 2012). Lammerhirt (2017) pointed out three critical barriers and the most important one is about findability: “data is hard to find.” Findability refers to the ease in which a user can access the datasets both from outside the website (through search engines and/or reference from other websites) and within the site.

The challenge of findability could be caused by various reasons. First, the challenge could be attributed to the availability of OGD. Some studies indicated that there was only processed data, no access to the original data (Janssen et al., 2012) and many types of data are not published (Zuiderwijk & Janssen, 2014). Also, OGD are fragmented because the OGD were published at numerous places on the internet (Van Veenstra & Van Den Broek, 2013; Zuiderwijk & Janssen, 2014; Zuiderwijk, Jeffery, & Janssen, 2012). Therefore, citizens have to check many different websites to find all the data they need (L. Koesten et al., 2017; Lammerhirt, 2017). Weerakkody et al. (2017) conducted a survey to learn citizens’ opinions about OGD usability, and they found that 107 people of the 516 surveyed stood by the fact that there is lack of clarity in the availability of open data.

Another possible reason that makes OGD hard to be found could be the technical difficulties, including interface design, system design and data retrieving technology (Roa et al., 2019). There are studies believed that OGD users could be deterred by the not user-friendly interface (Martin, 2014; Zuiderwijk and Jassen, 2014). Additionally, the data retrieval sys-

tem fails to support efficient searching and browsing for some reasons, such as no index or incomplete metadata (Janssen, Charalabidis and Ziuderwijk, 2012; Martin et al., 2013). Lammerhirt (2017) pointed out that users have to try many queries to access the needed dataset due to bad naming or website indexing. Also, Thomas et al. (2015) demonstrated that search results of datasets and their included tables that are provided by OGD portals are not ideal. It is hard for a user to know from a repository’s portal whether a useful dataset is available, and this problem is only likely to get worse. Zuiderwijk, Janssen, et al. (2012) claimed that even though the dataset is published on the OGD portal, users may not be able to find a certain dataset due to the primitive search capabilities of OGD portals. L. M. Koesten, Mayr, et al. (2019) led a discussion about online open data search in a workshop that was held in conjunction with SIGIR 2018. They came up with a list of technical challenges of searching data, including browsing and query support for structured and semi-structured data and learning to rank datasets.

The two foregoing reasons mentioned above discussing why people consider finding OGD difficult derive from the side of the data provider. Another reason could also derive from the side of users because some potential users may lack the necessary knowledge or skills to access data. For example, data users struggle to formulate or refine exact queries as they are unfamiliar to the government datasets (Liu & Jagadish, 2009; Swamiraj & Freund, 2015). Or, data users who are unable to acquire or employ the technical skills to access open data may be alienated from the OGD search and use (Gascó-Hernández et al., 2018; Janssen et al., 2012; C. Martin, 2014; Millette & Hosein, 2016).

Besides the above general challenges to access datasets, accessing OGD also has some unique difficulties. Users who want to access OGD can be an individual citizen who just wants to know more about the public work in her neighborhood, or a commercial startup company which aims to build real-time traffic alert Apps on mobile devices. The former user would mostly be a user with little technology capabilities or data literacy, whereas the latter user could be an expert on data manipulation or analysis. Both groups of users and all users in between should be supported in their access to OGD, which presents interesting, important and open challenges to the designers and managers of OGD repositories.

Gregory et al. (2019) claimed that improving accessibility of datasets is not an easy task.

They explain that to improve data discovery requires a data infrastructure and support systems. Also, understanding the user behaviors involved in data seeking behavior is equally essential, but “a user-focused, cross-disciplinary analysis of data retrieval practices is lacking” (Gregory et al., 2019; Ohno-Machado et al., 2017). Kacprzak et al. (2019) also mentioned that there are no systematic studies that investigate what properties of data consumer needs are essential for them to effectively search and discover datasets. In addition, from the perspective of data retrieval, the current web search engines are not well suited for searching datasets, since they are designed primarily for documents, not data (Cafarella et al., 2011; Kacprzak et al., 2017). Unlike online webpages, datasets contain various forms of data often without sufficient contextual information. Therefore, conventional content indexing methods used in web search engines do not always work for the datasets. Given this, although information retrieval (IR) has a long history, data retrieval is still a nascent field (Gregory et al., 2019). Therefore, the study of supporting the findability of OGD should be further conducted.

2.1.1.2 The Scholarly Work on Improving Findability of OGD. As shown above, there are extensive research efforts on examining the barriers of OGD project; correspondingly, many researchers have dedicated themselves to discover and provide solutions to solve the challenges. With regard to the challenge of findability, the corresponding scholarly work that contributes to addressing the difficulties can be classified into three main categories: studies concentrating on technical supports (i.e., data retrieval approaches and search interface design), studies regarding metadata quality, and studies focusing on OGD users, e.g. user OGD seeking behaviors and users’ needs of OGD.

Technical supports. As Kunze & Auer (2013) stated that unlike retrieving documents that are related to textual information description, dataset retrieval describes the process of returning relevant RDF datasets. Therefore, conventional content indexing methods used in web search engines do not always work for the datasets (L. Koesten et al., 2017; Swamiraj & Freund, 2015; Xiao et al., 2019). In 2009, Liu & Jagadish (2009) proposed an approach that tried to show all the relevant results first and then direct users to find what they need. They claimed that this framework is the first one that allows hierarchical database result browsing

and searching at the same time. After that, Kunze & Auer (2013) proposed an additional retrieval mechanism—dataset filtering. When querying, the entire set of available datasets is processed by a set of semantic filters, each of which is relevant to the query. The parameters of the semantic filters are set by the requester, which can clearly decide the relevance of a dataset to the given requirement. Another group of researchers also have similar concerns of poor searchability of current datasets portals (Thomas et al., 2015). They claimed that “the naive approach of full-text search is not appropriate.” They, therefore, came up with an alternative approach that attempted to use a ranking algorithm for tables to improve search results. In addition, to facilitate data search, Swamiraj & Freund (2015) developed an exploratory data search interface in particular for numerical datasets. They considered that the more novice search tasks of OGD are browsing or investigating rather than factual lookup or subject searching. Therefore, they created an exploratory data search interface to support novice users to learn and explore. Neumaier & Polleres (2019) also developed an interface design that allows spatial-temporal search of OGD.

Metadata quality. Data search systems on open data portals heavily rely on the text contained in metadata to facilitate keyword search. Therefore, the examination of metadata in OGD projects has been a topic of scholarly inquiry over the last few years (Kubler et al., 2018; Milic et al., 2018; Neumaier et al., 2016; Swamiraj & Freund, 2015; Xiao, Jeng, & He, 2018). Developing a metadata quality assessment framework is the first step in evaluating the quality of OGD portals’ metadata. Neumaier et al. (2016) designed a generic metadata quality assessment framework for various open data portals and applied it to monitor 260 portals with 1.1 M datasets. Kubler et al. (2018) also developed an open data portal quality framework (ODPQ) that allows end-users to easily and in real-time assess or rank open data portals, and this framework is also used to compare over 250 open data portals across 43 different countries. Both examinations that are based on the two frameworks revealed that metadata quality in OGD portals needs to be improved. In addition, there are suggestions concerning metadata: improving metadata structure and mapping identified metadata categories are conducive to disclosing relations between datasets in the same platform (Milic et al., 2018; Zuiderwijk, Jeffery, & Janssen, 2012); mapping OGD vocabulary contributes to cross search among all OGD portals, which makes data easy to be

searched (Binding & Tudhope, 2016); enhancing the quality of tags also serves to allow for effective findability of OGD (Tygel, 2016; Tygel et al., 2016).

OGD users. In terms of user studies, users’ OGD seeking behaviors and their needs have been identified, which helps data providers understand the OGD users’ needs and directs them towards better services. L. Koesten et al. (2017) employed an in-depth semi-structured interview method to understand user behaviors when interacting with OGD and then formulated a framework for human structured-data interaction. This framework discusses the processes of *task, search, evaluate, explore and use*. J. Crusoe & Ahlin (2019) proposed a process framework of user OGD interaction that contains user activities and related variations, where the included phases are: *start, identify, acquire, enrich and deploy*. These two frameworks have some similarities. For example, the *start* phase in J. Crusoe & Ahlin (2019)’s work is equivalent to the *task* phase in the L. Koesten et al. (2017)’s framework, which refers to the motivation of finding the desired data. Differences also exist; for example, the phases of *acquire and enrich* are not included in L. Koesten et al. (2017)’s work. In addition, there are studies using a transaction log analysis to study OGD user behaviors patterns, characteristics of issued queries, as well as data needs and additional contextual information (Ibáñez & Simperl, 2022; Kacprzak et al., 2019). For instance, Users desire to have additional filters that enable them to direct the search process and gain more desirable results, and temporal information is essential to a data search. (L. Koesten et al., 2017; Xiao et al., 2019). They also expect to identify the data links between the data internal and external of platforms to conveniently discover other available and pertinent datasets. (L. Koesten et al., 2017; Xiao et al., 2019). A feedback mechanism for users to report their issues (e.g., lack of the needed data or incomplete data), could empower the users to enhance the metadata and then facilitate data findability (J. Crusoe et al., 2019).

2.1.2 The Challenges and Ongoing Efforts of Understanding OGD

2.1.2.1 The Challenges of Understanding OGD. In addition to the challenge of findability, understandability is also one of the major difficulties for OGD users. J. Crusoe et al. (2019) found that compared to finding and accessing OGD, understanding OGD is

more difficult for users. Understandability refers to the ease with which data that are on a portal’s website can be read and interpreted by users. The data in open data platforms most often are available in raw data formats (Cranefield et al., 2014; Weerakkody et al., 2017) and users may be unfamiliar with definitions or categories that are adopted to present the data. Other main reasons that cause the difficulties of understanding OGD include lacking information to interpret data, lacking explanation for the meaning of data, and lacking knowledge to make sense of data (Eberhardt & Silveira, 2018; Graves & Hendler, 2013; Gurstein, 2011; Janssen et al., 2012; Nikiforova & McBride, 2021; Zuiderwijk, Janssen, et al., 2012). L. M. Koesten, Kacprzak, et al. (2019) indicated that the current descriptions about datasets do not necessarily increase understanding. If users cannot make sense of data, they would not be able to evaluate the data; let alone use the data they found.

Although the datasets come with certain metadata record that provides limited capabilities of text description, these structured descriptions often suffer problems, such as lacking detail descriptions, containing irrelevant elements for access, or missing certain essential elements (Janssen et al., 2012; Zuiderwijk, Janssen, et al., 2012; Zuiderwijk, Jeffery, & Janssen, 2012). Metadata is mainly created for collection management or long-term preservation, which often is not entirely associated with supporting access and making sense of data for a particular user. Therefore, comprehensive context information beyond metadata should be discovered to help users access and make sense of the OGD.

2.1.2.2 The Scholarly Work on Improving Understandability of OGD. Despite the fact that the OGD movement has flourished extensively, it is still in a relatively young stage, thus most of the current efforts on the OGD project are focused on facilitating data availability, accessibility and findability. In fact, the area of OGD understandability is essential but severely understudied. With a few research studies concerning improving understandability of OGD, the proposed supports can be classified into two main categories, which are more detailed information surrounding OGD and visualization tools. Existing literature suggests that guidance and help with understanding the content within OGD (e.g., descriptions of categories, what data is present, and so on) are important supports for augmenting the accessibility and understandability of OGD (Veljković et al., 2014; Verdegem & Verleye,

2009). For example, within the context of land use, Verburg et al. (2011) stated that it is essential to have clear and extended documentation to help users understand the data they are engaging with. L. Koesten et al. (2017) found out that more detailed information on how the original data was collected can aid users in a deeper understanding of the data and make a decision to trust in the data. Xiao et al. (2019) adopted mixed research methods approach to systematically explore and developed a framework of a dataguide which demonstrates the expected detailed information for assisting a user in making sense of the data, such as examples of how data has been used and the history of data formats (L. Koesten et al., 2017). These studies emphasized the importance of providing information and the context surrounding OGD.

Data visualization tools. Data visualization techniques has proven useful for understanding and communicating large amounts of data (Eberhardt & Silveira, 2018; Graves & Hendler, 2013). The data in open data platforms most often are available in raw data formats (Cranefield et al., 2014; Weerakkody et al., 2017), while visualizations allow users to easily discover trends and outliers that would be hard to detect based on the raw data and provides insights to users (Eberhardt & Silveira, 2018). Most used visualization technique is the map visualization, consisting of pointer map, route map, which present geospatial data (Radl et al., 2013), traffic data (Okamoto, 2017; Rocca et al., 2016) and other categories of data, e.g.city routing data. As described above, many potential users may be alienated from OGD by a lacking of expertise and technical knowledge to make sense of data. Leveraging these data visualization tools, especially, with the interactive visualization tools, would greatly alleviate this situation (Graves & Hendler, 2013; L. Koesten et al., 2017).

2.1.3 The Challenges and Ongoing Efforts of Using OGD

2.1.3.1 The Challenges of Using OGD. Besides the impediments to finding and understanding OGD, using data is a challenging barrier for OGD users in various aspects, including data availability, data quality, as well as ease in using OGD and the OGD platforms (Osagie et al., 2017; Saxena, 2018). Usability indicates the ease with which the data, platform interface, and functions can be easily used. Extant studies discovered that the avail-

ability of OGD is limited (J. Crusoe et al., 2019; Gascó-Hernández et al., 2018). The results show that the participants were frustrated by the lack of useful datasets. Participants have innovative ideas, but the ideas cannot be implemented due to the lack of pertinent OGD. In addition, data quality is one of the primary issues that impede users from using OGD. The issue of data quality consists of several facets, containing missing data values or errors (J. Crusoe et al., 2019; Osagie et al., 2017), lack of timeliness and legend information (Lee & Kwak, 2012; Saxena, 2018; Vetrò et al., 2016), and lack of OGD use standards (Ruijter & Meijer, 2020). These data quality obstacles can inhibit data from being used; even worse, the errors may lead to inaccurate analysis results.

Another two major challenges of using data are the ease of using OGD per se and its platforms. After the empirical examinations of OGD, some studies observed that the usability of OGD needs to be advanced. For example, the datasets lack consistency (J. Crusoe et al., 2019; E. G. Martin et al., 2017). Specifically, E. G. Martin et al. (2017) revealed that the provided data varies in size and scope. Some datasets contain demographic data, and some do not. Also, the lack of machine-readable formats (Janssen et al., 2012; Nikiforova & McBride, 2021) was raised by existing studies, indicating that the machine-readable formats can affect the use of datasets with large data sizes. In addition, J. Crusoe et al. (2019) identified the user challenge of aggregating and transforming data. The difficulties were caused because the datasets may use different metrics, have no unique identifiers, and the naming terminology varies. For instance, with the same neighborhoods, the neighborhood names were inconsistent in different datasets. In the meantime, the complex OGD platform design (Osagie et al., 2017) and the lack of interactive functionalities (Gascó-Hernández et al., 2018; Zuiderwijk et al., 2016) were also considered crucial factors limiting data use. As described by Osagie et al. (2017), the complexity of the system generates difficulties in using data for non-expert users.

Other challenges concerning the platform side were found by existing research studies, including the lack of effective communication between OGD users and data providers (Gascó-Hernández et al., 2018; Saxena, 2018), the lack of guidance or tutorials about the portal usage (Li et al., 2022), and the lack of trust in OGD (Saxena, 2018).

The discussions above primarily focus on the challenges caused by the OGD portal side.

From the user side, extant studies also discovered an impediment to using OGD, which is the requirement of knowing necessary knowledge and data literacy skills (J. Crusoe et al., 2019; Liu & Jagadish, 2009; C. Martin, 2014; Xiao et al., 2022). For instance, J. Crusoe et al. (2019) found that data needs special knowledge to understand, including domain knowledge, the knowledge of preparing data, and the abilities of forward-thinking and backtracking. Based on an observational study, Li et al. (2022) observed that low data literacy is a significant obstacle to using data. Extant studies point out that most potential OGD users lack OGD literacy skills (Boychuk et al., 2016; Gascó-Hernández et al., 2018; Li et al., 2022), especially computational skills (Xiao et al., 2022). Even though OGD portals can offer interactive tools, it is impossible to fulfill all users' needs when a user is without OGD literacy.

2.1.3.2 The Scholarly Work on Improving Usability of OGD. With the challenges of using data, studies began to find solutions to address these difficulties from various dimensions. In fact, the scholarly works on improving the findability and understandability of OGD, e.g., enhancing metadata quality and creating visualization tools, are also for advancing the usability of OGD. In addition, effectively examining the performance of OGD projects, improving data quality, providing technical support, and promoting OGD trust are argued as solutions to improve OGD use by existing research studies. Dawes et al. (2016) developed an ecosystem model for facilitating the plan and design of OGD projects. This model aims to be useful for both design and evaluation of open government data programs. Existing studies also examined OGD portals' performances through various analysis methods, such as a longitudinal cross-sector analysis (Chatfield & Reddick, 2017) or a comparative cross-jurisdictional analysis (Kassen, 2018). As for the issue of data quality, studies explored the evaluation framework for examining data quality concentrating on OGD (Dorobăț & Posea, 2021; Kučera et al., 2013; Vetrò et al., 2016). For example, a metric-based evaluation framework to assess the quality of OGD (Vetrò et al., 2016).

Regarding the development of technical support, Ermilov et al. (2013) designed an application for transforming statistical data and facets for spatial data, enabling users to perform simple data mining tasks on the transformed tabular data. Another study developed an approach that is lifting the often semantically shallow datasets registered at OGD portals to

Linked Data in order to make data portals the center of a distributed global data warehouse (Waal et al., 2014). Abbas & Ojo (2013) proposed the idea of taking advantage of linked data to enable global access to spatial data, and they developed a Reference Architecture for building interoperable Linked spatial data infrastructures. Given that OGD involves a great part of geospatial data and corresponding analysis, there are studies focusing on interactive Geo-Visualization designs to improve data use and augment citizens' engagement (Degbelo, 2022; Fechner & Kray, 2014).

Regarding the enhancement of the trustability of OGD, previous research studies shed light on the factors associated with the citizens' trust in the OGD, which reflected that citizens do not easily trust the data. To increase the trustworthiness of OGD, Purwanto et al. (2020) empirically investigated the attributes that can affect citizens' confidence in OGD. They found that system quality and service quality can influence citizens' perspectives of OGD. Another study proposed that when the data is opened to the public, and the results of data reuse are replicated, it can contribute to the trust of citizens (Meijer et al., 2014).

At last, the difficulty of using data from the user side concerning OGD literacy was discussed. Researchers acknowledged this critical factor. Wolff, Montaner, & Kortuem (2016) claimed that "as open data becomes established as part of everyday life, the ability for the average citizen to have some level of data literacy is increasingly important." They developed an approach to teaching data skills in schools based on the principles of narrative and inquiry-based learning. Some studies discussed the significance of data literacy (Frank et al., 2016; Wolff, Gooch, et al., 2016) and mentioned the possible design for data literacy by including human-centered approaches (Bhargava et al., 2015). By conducting a literature review, D. Crusoe (2016) discussed the definition of OGD literacy, and the skills and knowledge that are attributed to OGD literacy, e.g., "knowledge of what data are, how they are collected, analyzed, visualized and shared, and the understanding of how data are applied for benefit or detriment." Based on this literature review, few research studies worked on OGD literacy capabilities, which inspired me to investigate comprehensive and fundamental OGD literacy capacities from a user-centered perspective.

Section 2.1 summarizes the identified challenges of finding, understanding, and using OGD and the proposed solutions or suggestions for ameliorating users' difficulties. This

review offers me a holistic and multifarious picture of OGD today’s situation. Also, in terms of this examination, the research method of semi-structured interviews is the most used approach to explore user behaviors of OGD. Transaction log analysis has also been adopted to look into the characteristics of issued queries and the patterns of data seeking behaviors. Therefore, the two methods proven to be successful in investigating user behaviors in the field of human OGD interaction. Also, despite taking advantage of social group data to observe and analyze users’ challenges and behaviors has yet to be identified in OGD field, it is an effective method in health information behavior exploration. Therefore, I decided to use the content analysis method to analyze a local-level OGD portal’s user group data as well, which brings a new perspective to this field.

2.2 Human Information Interaction

The new concept of human data interaction (HDI) stems from the fields of data visualization and human-computer interaction (HCI) (Elmqvist, 2011; Trajkova et al., 2020). However, Mortier et al. (2013) considered HDI to be inherently interdisciplinary; it is not only attributed to traditional computer science fields such as data processing, but also to be discussed in the disciplines of psychology, behavioral economics, and sociology. By reviewing a number of scholarly works concerning HDI, Victorelli et al. (2020) argued that there is a lack of literature that focuses on HDI in designing information systems. They proposed a set of recommendations for HDI in other specific domains, such as health informatics, urbanism, and smart cities. Therefore, I aspire to conceive HDI from the lens of Information Science, and the most popular concept corresponding to HDI is human information interaction (HII). Kunze & Auer (2013) propose that “dataset retrieval is a specialization of information retrieval.” L. Koesten et al. (2017) also considered that data seeking behavior is information seeking behavior but with new sources of data. This section reviews the concepts of HII and the corresponding models.

2.2.1 The Concepts of Human Information Interaction

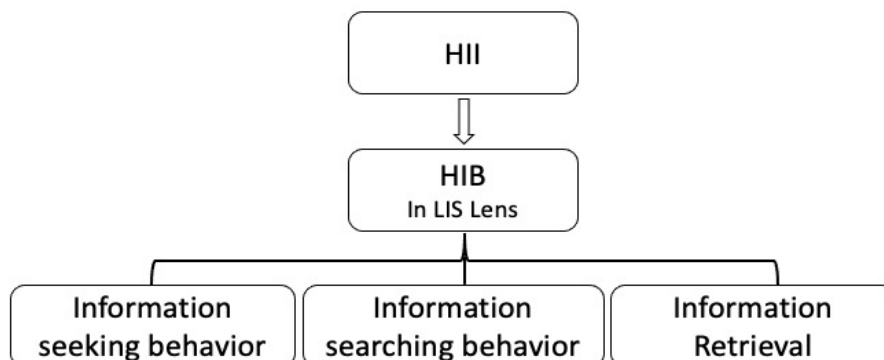


Figure 4: The Core Concepts of HII

HII is defined as an area of research that investigates how humans interact with information. It is manifested as **human information behavior (HIB)** in library and information science and as HCI in computer science (Fidel, 2012, p.274). The concept of **HIB** is “the totality of human behavior in relation to sources and channels of information, including both active and passive information seeking, and information use” (Wilson, 2000). Ford (2015) proposed that “HIB is all about how we need, find, process, and use information.” **HIB** encompasses **information-seeking behaviors**, **information retrieval**, using information, filtering information, avoiding information, organizing information and representing information (Fidel, 2012, p.21-43). Within the LIS field, **information seeking behavior** and **information retrieval** are the two research areas that have already formulated a set of agreed upon concepts (Fidel, 2012, p.21). **Information seeking behavior** studies the strategies people use to look for information (e.g., browsing and searching) and probably embrace the selection and use a variety of search tools (p.21 Fidel, 2012; Ford, 2015, p.14). The objective of **Information retrieval** is to “build and investigate retrieval models and mechanisms for computer-based systems that retrieve information in response to user request.” (Fidel, 2012, p.21). The studies on information seeking behavior are more from the people’s side, while information retrieval has been discovering the system of design for assisting in information search. Another critical concept in HIB is **information searching behavior** that is recognized as the behaviors of using particular search tools, such as a search engine or

database to acquire information (Ford, 2015, p.14). The concepts of **information seeking behavior** and **information searching behavior** were distinguished by Wilson (1999). He argued that “[i]nformation Searching Behavior is the ‘micro-level’ of behavior employed by the searcher in interacting with information systems of all kinds.” Wilson (1999) also considered information searching behavior as a sub-set of information seeking behavior in his nested model, while Fidel (2012) used the two concepts interchangeably according to the context in which they are addressed. Figure 4 presents the core concepts in the field of HII.

2.2.2 Two HIB models: Big Six and Information Search Process

I present two models of HIB in this section, which are Eisenberg & Berkowitz (1990)’s the Big Six and Kuhlthau (1991)’s model of the information search process. HII and HIB have a long history, and accordingly, a number of HIB models have been developed and applied into various settings. The rationales of choosing to demonstrate these two models are: 1) both of the models are action models that represent activities during the process of interacting with information. To explore OGD user behaviors of interacting with data, their activities must be observed. 2) Most HIB models depict the process focusing on seeking behaviors, while Eisenberg & Berkowitz (1990)’s model represents an entire process of information interacting, covering the seeking and synthesizing processes. 3) Kuhlthau (1991)’s model explicitly and specifically demonstrates the process of seeking information, including general information seeking and focused information seeking. This model has been applied in various themes, especially, as L. Koesten et al. (2017) constructed the framework for human structure-data interaction by referencing this model.

Eisenberg & Berkowitz (1990)’s The Big Six. The Big Six is an action model which contains six successive steps leading people to solve an information problem or make a decision that is based on information. The six steps are *task definition*, *information-seeking strategies*, *location and access*, *use of information*, *synthesis*, and *evaluation*. Specifically, *task definition* refers to defining the problem and identifying the information needs. *Information seeking strategies* indicates all possible source determination and the best source selection. *Location and access* points out the behaviors of locating sources and finding information

within sources. The *use of information* signifies engaging, such as reading and viewing, and extracting relevant information. *Synthesis* refers to the organizing of information from multiple sources and the presenting of the information. The last step, *evaluation*, is to judge the result and the process.

This model originated from the education of information literacy in elementary school. Now, it has been expanded to other education settings and to the themes whenever people are encountering information difficulties. Eisenberg & Berkowitz (1990) conceive this model as not only a set of skills that help people to tackle information issues but also an approach that guides people to learn the information problem-solving process.

Kuhlthau's model of the information search process (ISP). Kuhlthau (1991)'s model is one of the most prevalent information seeking models in LIS area. This model was developed based on a series of five empirical studies that investigated user common experiences when seeking information, and it focuses on users' perspectives. Six stages are contained in this model: *initiation, selection, exploration, formulation, collection, and presentation*. At the *initiation* stage, an individual realizes an information need. Then, in the section stage (*selection*), she begins to identify and select the general topic to be investigated and the approach to be used. The typical actions include the communications with others or searching for relevant information and then conducting a cursory review of alternative topics. The goal in the third stage, *exploration*, is to extend personal understanding. The specific actions consist of locating information about the general topic, being informed in terms of reading, and establishing connections between the new information to your known information. The fourth stage is *formulation*, that is to form a focus based on the information that was reviewed, and then the person starts to gather information related to the focused topic in the *collection (the fifth)* stage. The last stage is *presentation*, which is to finalize the search and prepare to present or use the search results.

In this section, I summarized each stage of Kuhlthau's model by concentrating on the realm of actions. In fact, as shown in Figure 5, the novel contribution of this model to the field of information seeking behavior is the three realms: the affective (feelings), the cognitive (thoughts), and the physical (actions). The three facets holistically illustrate people's experiences when looking for information and the transformation of the facets within the

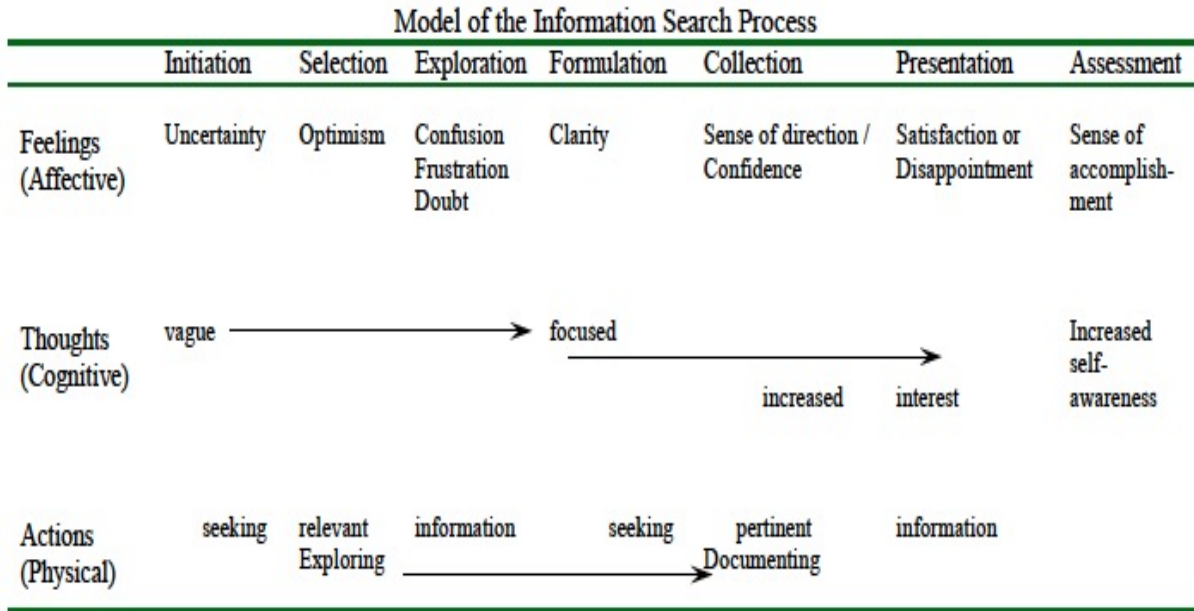


Figure 5: Kuhlthau's Model of the Information Search Process (ISP) (Kuhlthau, 1991)

search process is explored and presented. This model is the first model that considers the role of feeling in the seeking behaviors. Also, this model has been applied to many areas and cited thousands of times.

2.2.3 Information Needs and Search Strategy

The two models offer a sense of understanding for HIB models, and they also raise two seminal concepts within the activities of HIB, which are **information needs** and **information seeking strategies**. Fidel (2012) claimed that the concept of information needs is a foundation on all the processes of in HII, since all search decisions and activities are determined by information needs. However, the definition of information needs has been discussed for several decades, and there is no agreed-upon conclusion.

Taylor (1968) proposed an understanding for information needs, which is the earliest and most discussed depiction of information needs. The following presents the four levels of information needs with the consecutive cognitive stages.

level 1 - the actual, but unexpressed need for information (the visceral need);

level 2 - the conscious, within-brain description of the need (the conscious need);
level 3 - the formal statement of the need (the formalized need);
level 4 - the questions as presented to the information system (the compromised need).

In the *visceral* stage, the inquirer merely feels a sense of need for information, the needs are unclear, vaguely recognized, and could be varied. In the *conscious* stage, people could identify that there is a need for information, but it is still ill-defined. However, the inquirer may want to continue to develop it, such as by means of talking to others. Then, he may be able to form a qualified and rational statement of his question, which enters to the stage of *formalized* need. In the *compromised* stage, the formalized questions/information need are submitted to the information system.

Another early concept of information needs was described by Ford (1980) as “an awareness of a state of ‘not knowing’ or some conceptual incongruity in which the learner’s ‘cognitive structure is not adequate to the task’.”

The more recent research works have considered information needs as the prelude to a search process. For example, information need motivates information seeking behavior (Bruce, 2005). Whereas, the latest view indicates that there is a shift from *information need* to *task*, which means task stimulates searching for information (Fidel, 2012, p.84). Based on a cursory literature review, Byström & Järvelin (1995) concluded that “it has been often accepted that information needs and the information-seeking processes depend on worker’s tasks.” He argued that “information needs” is the gap between the worker’s knowledge about the task and the perceived requirements of the tasks.

Compared to *information needs*, which are relatively stable, *search strategy* is to solve the dynamic part of the search process. Bates (1981) defined *search strategies* as “an approach to or plan for a whole search. A search strategy is used to inform or to determine specific search formulation decisions; it operates at a level above term choice and command use.” Marchionini (1997) argued that “A strategy is the approach that an information seeker takes to a problem. Strategies are those sets of ordered tactics that are consciously selected, applied, and monitored to solve an information problem.” These two concepts are similar to each other, which are under the broader definition that is the variety of behaviors in which people engage when searching for information. These strategies can be conceived as

interactions between people and other components of the IR system (Belkin et al., 1995, 1993). The search strategies can be classified into the browsing strategy, the analytical strategy, the empirical strategy, the Know-Site strategy, and the similarity strategy (Fidel, 2012).

2.3 Theories and Models on Sensemaking

Sensemaking is a term understood as the processes through which people interpret and give meaning to their experiences (LAM et al., 2016), and it has been used in a number of disciplines, such as HIB, organization communication, knowledge management, cognitive system engineer, and HCI (LAM et al., 2016). The notion of sensemaking in the field of HCI in early 1990s was framed as the process of searching for, forming and working with meaningful representations in order to facilitate insight and subsequent intelligent action (Pirolli & Russell, 2011; Russell et al., 1993). In this dissertation, I adopt the understanding of sensemaking from Pirolli & Russell (2011), that delineated that “sensemaking involves not only finding information but also requires learning about new domains, solving ill-structured problems, acquiring situation awareness, and participating in social exchanges of knowledge.” This definition suggests an active processing of information to achieve understanding. In this section, I review two models of sensemaking, one is Dervin (1999)’s sensemaking theory; the other is Pirolli & Card (2005)’s sensemaking model, which helps formulate the proposed conceptual model of HDI in this dissertation.

2.3.1 Dervin’s Sensemaking Theory

Dervin’s sensemaking theory is one of the most popular theories in LIS area. The first skeleton of this theory was formed in 1972, and it was first illustrated as sensemaking in 1983 (Dervin, 2005). Until today, Dervin’s sensemaking theory continues to be enriched and expanded its conceptual construct (Dervin, 2005; Fidel, 2012). It was proposed as a theory for information design (Dervin, 1999) and it has been applied to numerous areas as

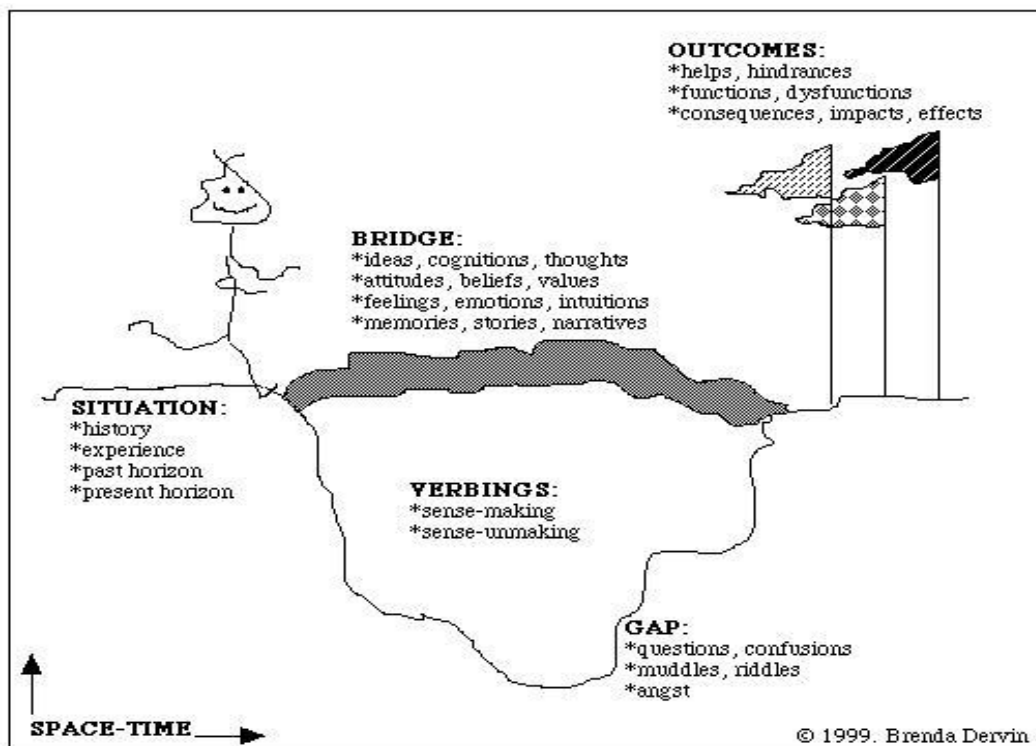


Figure 6: Dervin's Sensemaking Theory (Dervin, 2005)

a methodological guidance, such as libraries, media systems, educational institutions, and health care delivery (Dervin et al., 2012). Figure 6 is a graphic presentation of Dervin’s sensemaking methodology.

Dervin et al. (2012) proposed that “Sensemaking methodology rests on a metaphor of human movement through a time-space that is discontinuous always at least in part; and how humans make and unmake sense as they journey through these movements”(Dervin et al., 2012). Dervin (1999) perceives that “Sensemaking metaphor must be understood as a highly abstract framework.” Figure 6 is a metaphor of a sensemaking process, which describes a human movement process. An individual aims to achieve a goal. First, this person needs to be embedded in a certain dynamic context and situation, including the knowledge, skills, and experience that she already owns. Still, there is something she may be confused about or lack the relevant understanding or skills, and then she is stopped by the gaps. Consequently, to move to a new situation, she needs to build the bridge by taking advantage of various sources of information, and ultimately reach the goal. This is a simplified narrative of Dervin’s sensemaking methodology. Sensemaking relies heavily on concepts of time, space, movement and gap, and it is not merely a purposive, linear, problem-solving activity. Also, the sensemaking triangle: situation, gap/bridge, and outcomes are highlighted by Dervin (1999). She stressed that one needs to identify the gap (i.e., gap finding) and make use of the corresponding bridge to achieve the outcome of moving to the other end of the gap via the bridge (i.e., gap bridging) and finally earn outcomes. This sensemaking methodology intended to guide research (e.g., question formation, data collection, data analysis), and dialogue (e.g., organizational, societal, digital).

2.3.2 Pirrolli & Card’s Conceptual Model of sensemaking

As presented in Figure 7, Pirolli & Card (2005) perceived sensemaking processes from two major loops of activities: a *foraging loop* and a *sense making loop*. The activities in the foraging loop consist of seeking, searching and filtering information, as well as reading and extracting information. The actions in a sensemaking loop covers iterative development of a mental model from the schema that best fits the evidence, including extracting, reorga-

nizing and reevaluating information. This processing can be driven by bottom-up processes, top-down process, and lots of back loops. This flow demonstrates the transformation of information from raw information to a presentation.

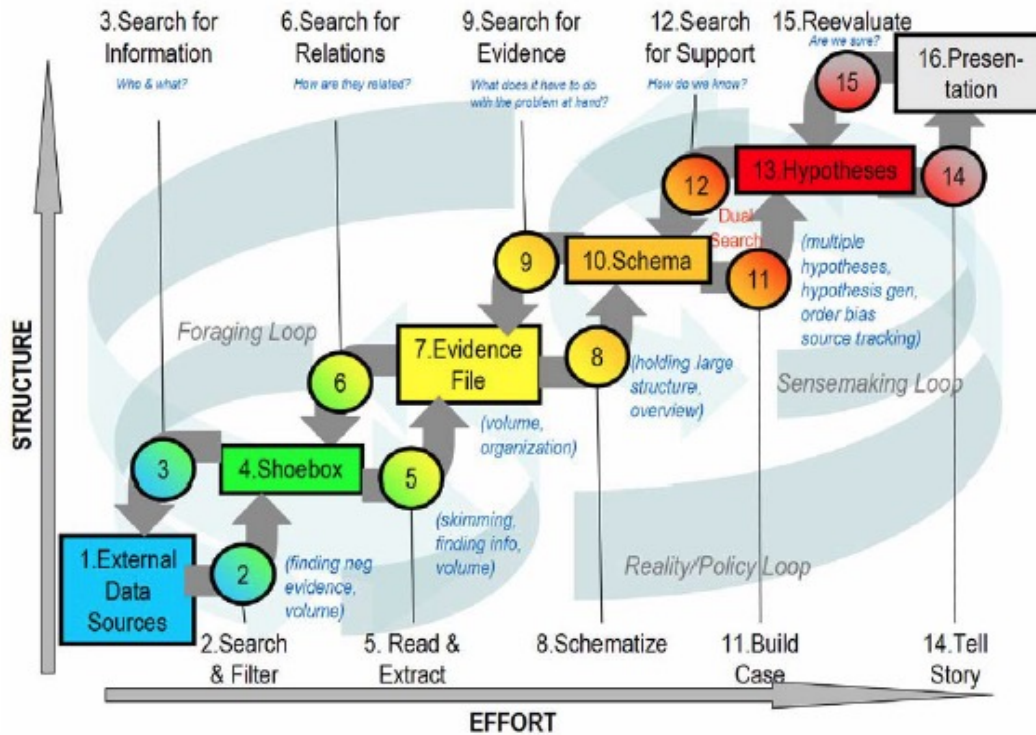


Figure 7: Pirrolli & Card's Conceptual Model of Sensemaking in Intelligence Analysis (Pirrolli & Card, 2005)

As presented in Figure 7, Pirrolli & Card (2005) elaborated on all the terms and processes. The rectangular boxes represent an approximate data flow. The circles represent the process flow. *External Data Sources* refers to all the raw evidence in all sources that an inquirer reaches. *Shoebox* is the relevant evidence for later processing that is selected from *External Data Sources*, which is a smaller subset of the external data. *Evidence file* refers to more relevant information extracted from the items in the *Shoebox*. *Schemes* indicates the re-representation of the collected information that is organized or shaped to support insight or to be used more easily to draw conclusions. *Hypotheses* are the tentative representation of the conclusions with supporting arguments. *Presentation* refers to a work product or a presentation that manifest the goal of the entire process.

* The Bottom-up process

Search and filter. The activities include an inquirer searching for information and filtering (judging) information for relevance in external data sources, such as Pub-med or databases. Then, the inquirer collects the relevant documents into the *Shoebox*.

Read and extract. The activities include the inquirer reading to extract more relevant snippets from the *Shoebox*. The evidence extracted at this stage may trigger new hypotheses and searches.

Schematize. The activities include the inquirer forming the collected information in some schematic way, but it may be a simple or informal way. For example, the information collected from the *read and extract* stage may be organized into small-scale stories.

Build case. The activities include the inquirer building a theory or case to support or disconfirm hypotheses after forming a simple schematic story.

Tell story. The inquirer makes a presentation or a work product for some audience (client).

* The Top-down Process

Re-evaluate refers to the activities that ensue when the inquiries or feedback from audiences of a presentation requires a re-evaluation of the current theories or hypotheses.

Search for support means that the re-evaluation or analysis requires to a review of the lower-level schematic organization of basic facts.

Search for evidence denotes that the re-evaluation or analysis requires reviewing the collected evidence or the searches for new evidence.

Search for relations discusses that new information may generate new patterns that may generate new hypotheses that may generate new searches and data extraction for the *Shoebox*.

Search for information emerges when the hypotheses lead to a deeper dig into the raw data.

This conceptual model encompasses user seeking behaviors and sensemaking behaviors. Therefore, I decided to use it as a component to design the conceptual model of HDI for this dissertation project

2.4 Human Data Interaction & HDI in Open Government Data

As Trajkova et al. (2020) stated that the term HDI was used for the first time in a research paper of data visualization that discusses the analysis of multi-variate data. To the best of my knowledge, the first definition of HDI was given by Elmqvist (2011). He defined HDI as “the human manipulation, analysis, and sensemaking of large, unstructured, and complex datasets.” This definition has been discussed in HCI literature, such as the works of Hornung et al. (2015) and Trajkova et al. (2020). In another perspective, Mortier et al. (2013) proposed that HDI concerns interaction generally between humans, datasets, and analytics. They particularly stressed that HDI aims to understand not only the raw or derived data out there but also how and by whom they are used, as well as how people might desire and act to influence and ideally benefit from the data and their use. Since HDI is still an emerging field, only two frameworks are discussing HDI in the OGD field. One is the *Framework for Human Structured-data Interaction* proposed by L. Koesten et al. (2017) and the other one is the *Process Framework for users’ activities for using OGD* proposed by J. Crusoe & Ahlin (2019).

2.4.1 The Framework for Human Structured-data Interaction

TASK	SEARCH	EVALUATE	EXPLORE	USE
goal or process oriented	STRATEGIES: <ul style="list-style-type: none">• web search engines• data portals• asking people• FOI request	<ul style="list-style-type: none">• relevance¹• usability²• quality³	<ul style="list-style-type: none">• basic visual scan• obvious errors• summarizing statistics• headers• documentation• metadata	
linking				
time series analysis				
summarising				
presenting				
exporting				
1: context, coverage, summary, original purpose, granularity, time frame	2: labeling, documentation, licence, access, machine readability, language used, format, ability to share, format	3: collection methods, provenance, consistency of formatting/labeling, completeness, what's been excluded		

Figure 8: Framework for Interacting with Structured Data. (L. Koesten et al., 2017)

L. Koesten et al. (2017)’s HDI framework focuses on users’ online OGD seeking behavior by referencing Kuhlthau (1991)’s HIB model, containing task context, data search, data relevance evaluation, explore, and use (see Figure 8).

In this study, the *task* is recognized as the activities of using data as a material to create something, such as a service or a tool. The *task* is classified into two broad categories: “(1) Process-oriented task - people think of these tasks as doing something transformative with data, and (2) Goal-oriented task - people think of data as a means to an end” (L. Koesten et al., 2017). The researchers also proposed another way to categorize tasks into five types: (1) *Linking*, which refers to finding commonalities and differences between two or more datasets (2) *Time series analysis*, which signifies the task that identifies trends or detects and predicts events using the data ordered by time (3) *Summarizing* denotes the activities of creating a more compact, meaningful representation of the data (4) *Presenting* involves the activities of transforming data into human-friendly formats, such as visualizing them (5) *Exporting* refers to the activities of producing and publishing a dataset in a given format, e.g., metadata.

Search describes how people search for datasets. The major methods are searching on the web and searching on portals. In terms of L. Koesten et al. (2017)’s study, most participants adopt Google to find data online. Another two approaches to look for datasets are asking people (colleagues and professionals) and directly asking the data publishers.

Evaluation and Exploration. After identifying the dataset of interest, users experience two broad stages: evaluation and exploration. Evaluation is described as *the first look*, users decide whether or not to use the dataset by utilizing three criteria: relevance to the task, usability, and quality. Then, users begin to build a notion of quality and trust of the data during the *exploration* stage.

Although this framework is called human structured-data interaction, it concentrates on the processes of seeking behaviors, excluding the sensemaking behaviors. Therefore, I combine this model with Pirolli & Card (2005) sensemaking model to formulate a HDI conceptual model.

2.4.2 Process Framework for Users' Activities for Using OGD

After L. Koesten et al. (2017)'s framework on HDI, J. Crusoe & Ahlin (2019) examined OGD user activities when interacting with data and came up with a process framework of OGD user activities and the corresponding variations. As presented in Figure 9, this framework covers five broader stages with nine specific processes.

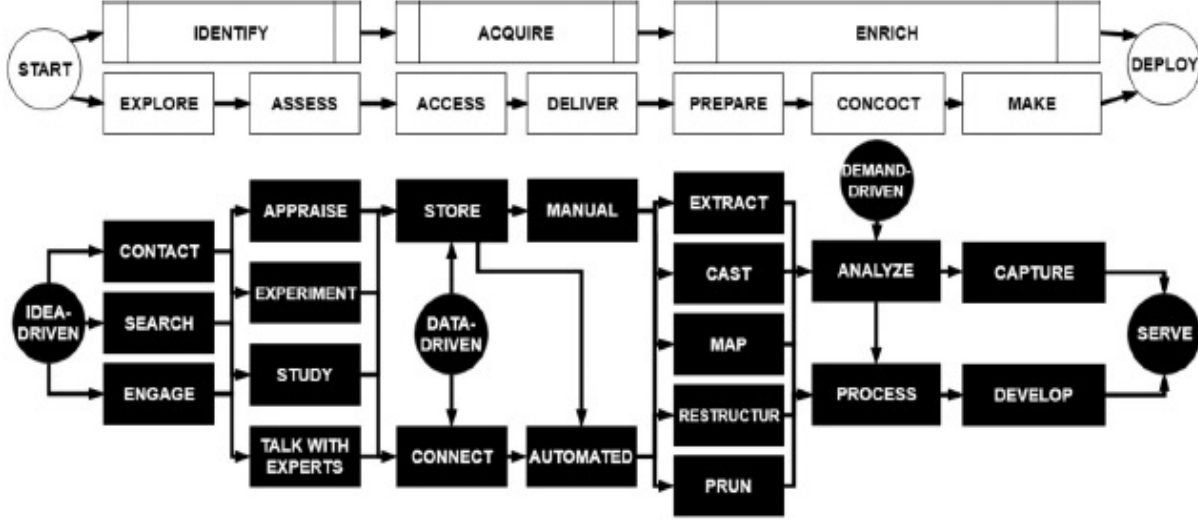


Figure 9: The User Process with Activities and Variations. (J. Crusoe & Ahlin, 2019)

In the *start* stage, the user starts the interaction with the context of idea-driven, data-driven, or demand-driven. This stage is similar to L. Koesten et al. (2017)'s *task* stage, which sets the motivation for interacting with data.

The *identify* stage includes the activities of exploring for and assessing OGD. Specifically, users may contact, search, and engage with the datasets and assess whether the data is the right data for the intended use. This process involves data quality evaluation and surface-level data interpretations. This stage is corresponding to the *evaluation* stage in L. Koesten et al. (2017)'s work.

The *acquire* stage describes accessing data, including automated accessing data or a need for publishers to deliver the needed data to users manually. It emerges when users have identified the data of interest, directly downloaded the data from the website, or acquired the data from the data publisher to manually deliver to them. This stage is not included in

L. Koesten et al. (2017)’s framework.

The *enrich* stage consists of activities of *prepare*, *concoct*, and *make*. *Prepare* includes the variations extract, cast, map, restructure, and prune. The *concoct* activity refers to combing the various datasets to analyze, process, and interpret the data. *Make* is to understand the intended end-user and the intended use to help users capture and develop the data. This stage is not covered by L. Koesten et al. (2017)’s framework either.

The *deployment* stage is the end of the user process, which refers to the product or service that can serve external end-users or be used for personal gains.

This framework provides a detailed process of interacting with OGD, including online and offline behaviors. I decided to employ L. Koesten et al. (2017)’s framework as a component of the proposed conceptual model of HDI mainly because this framework focuses on users’ online data-seeking behavior, and they studied it within the context of OGD.

2.4.3 HDI in Open Government Data

The proliferation of open data movements and ubiquitous computing technologies of recent years bring with it a dramatic increase in data collection and data generation. Meanwhile, governments and public sectors enthusiastically encourage the public to engaged with these data to create values. Aligning with this trend, HDI is emerging in the field of OGD. To the best of my knowledge, there are three papers in the field of OGD that mentioned or thoroughly discussed concepts of HDI, which may not use the exact term, *Human Data Interaction*, but are affiliated with it. The concept of HDI was applied for the first time in OGD by L. Koesten et al. (2017) to refer to a framework for human structured-data interaction, which straightforwardly discussed the interaction between humans and OGD. It was called *human structured-data interaction*, as most of OGD are structured data. The details of this framework are introduced in the last section. Xiao et al. (2019) adopted this HDI framework to develop the research instrument for their study to explore OGD users’ difficulties in each stage when interacting with OGD. Another study formulated a user process framework of engaging with OGD (J. Crusoe & Ahlin, 2019). Even though this framework was not directly called *HDI*, it describes the process of users working with OGD, including

the activities and variations. Compared to L. Koesten et al. (2017)’s work, in addition to the stages of *task*, *search*, *evaluation*, *explore*, J. Crusoe & Ahlin (2019)’s framework also contains an *enrich phase*, which consists of the activities of preparing, concocting and making, specifically, the activities of data extraction, data casting, data mapping and so on. J. Crusoe & Ahlin (2019) claimed that the significant contribution of this user process is that it can act as a foundation for future research about OGD use, be a first step towards the creation of a descriptive theory for the usage of OGD, as well as inspire the understandings of the different strategies users use and how they can best be supported.

2.5 The Development of Data Literacy

Before addressing data literacy (DL), we have to briefly discuss information literacy (IL) first because DL is an extension of IL. The concept of IL has been researched and developed for several decades. In 1974, the concept of “information literacy” was introduced for the first time by Paul Zurkowski, who was the president of the Information Industry Association. Zurkowski brought up “information literacy” in a proposal submitted to the National Commission on Libraries and Information Science (Behrens, 1994; Spitzer et al., 1998). He believed that “People trained in the application of information resources to their work can be called information literates. They have learned techniques and skills for utilizing the wide range of information tools as well as primary sources in molding information- solutions to their problems” (Spitzer et al., 1998). With the development of the concepts of information and information technologies, scholars have begun to propose their own understandings of information literacy. For example, as seen in Stephenson & Caravello (2007)’s work, data literacy is a new concept recognized by many scholars as a necessary component of information literacy. Given that more and more people can access data through open data initiatives, data literacy has obtained more and more attention in different fields, including doing research or understanding the news or government policies. Therefore, the definition of data literacy is often varied.

DL is frequently connected with statistical literacy (SL) and has further been developed

by social science and open data communities (Prado & Marzal, 2013). According to the literature on the topic, the evolution of DL has experienced three main phases. In the first phase, the term DL appeared in academic papers written in 2004 Shields (2005) and Hunt (2005). Both of them discussed the needs for DL, as well as the connections and differences amongst IL, SL and DL. Then, similarly to IL, with the development of the technologies, networks, various software and applications, the major efforts to define DL often embraced identifying, understanding, operating on, using and managing data. In the first two developmental phases, the definitions of DL were more focused on defining the skills associated with students in higher education, researchers, and scientists. The third phase points in a future direction. In recent years, the open data movement has led to the idea of a data literate citizenry (Twidale et al., 2013), which would lead to some different requirements to be data literate. This section illustrates the evolution of DL in detail based on the literature.

2.5.1 The Beginning Phase (An Information Literacy and Statistical Literacy Phase)

The demands for DL started within the social science and business disciplines. Shields (2005) pointed out that the students in majors that require data analysis or statistics needed a class on data literacy. During that time, “how to use and assess research” is one of the most essential elements for a sociology curriculum (Wagenaar, 2004). Additionally, Shields (2005) proposed that understanding and using basic statistical concepts was crucial to discuss current social issues. Hence, he considered SL was a significant skill to social science students and argued that to be statistically literate, “they must be able to think critically about basic descriptive statistics, analyzing, interpreting and evaluating statistics as evidence is a special skill.” On the other hand, Shields (2005) identified DL as being able to “access, manipulate and summarize the data.” He also claimed that the way of data being collected, converted and manipulated can greatly influence a numerical value of a statistic. In the same year, Hunt (2005) also discussed DL, SL and IL together, but he proposed a different idea. First, he thought SL, quantitative reasoning or quantitative

literacy, numeracy and data literacy all have roughly the same meaning. Second, Hunt (2005) considered implementing DL is very different from IL. He provided an example from a course related to IL and DL. He noted that he can assume students in this course know how to read, and use a browser to find information, but he is not able to assume that students know how to use a spreadsheet to process data. This concern conflicts with one of Shields (2005)'s arguments that an information literate must be a statistical literate. In 2007, Stephenson & Caravello (2007) conducted a pilot project about incorporating DL into undergraduate IL programs in the social sciences. The DL modules they developed were engaging students in activities to help them effectively use statistical resources in course assignments/papers and critically evaluate graphical data. The DL module provided minimal introduction about basic statistical concepts, such as proportion and rate, that would assist students to read tables and critically think about the information in the tables. For future plans, they considered developing DL as part of IL.

2.5.2 The Second Phase (A Technical and Skill-based Phase)

In 2010, Qin & D'Ignazio (2010) proposed the term science data literacy (SDL) that emphasizes "the ability to understand, use, and manage science data." SDL is more focused on a "functional ability in data collection, processing, management, evaluation and use." This definition considers two groups of data users —e-science data literates and e-science data management professionals. Also, this paper did not distinguish between DL and SL, but drew a table to display the differences of IL, SDL and digital literacy. This preferred definition combined the capabilities of data collection and data management into DL. Carlson & Johnston (2015) have a similar idea that merges the concepts of researcher-as- producer and researcher-as consumer of data products. They also awakened a new term, "data information literacy," to describe this idea. Next, Carlson & Johnston (2015) defined DL as the ability to process, sort, and filter vast quantities of information, which requires knowing how to search, how to filter and process, and to produce and synthesize. He pointed out that the current problem is the vast sea of textual, audio, and video data that we wade through every day, so the new skill of helping filter and sort through this information is necessary.

Through this, the new elements of search and filter were proposed to DL.

Prado & Marzal (2013) reviewed six standards that are relevant to DL and summarized the competencies explicitly associated with DL. I also reviewed the DL competencies that were discovered in the studies of Mandinach & Gummer (2013), ODI (2015), and D'Ignazio & Bhargava (2016), and combined them with Prado & Marzal (2013)'s work. A synthesis is shown as follows:

- 1) Ability to identify the context in which data are produced and reused (data lifecycle)
- 2) Ability to recognize source data value, types and formats
- 3) Ability to determine when data are needed
- 4) Ability to access data sources appropriate to the information needed
- 5) Ability to critically assess data and their sources
- 6) Ability to determine and use suitable research methods
- 7) Ability to handle and analyze data
- 8) Knowing how to select and synthesize data and combine them with other information sources and prior knowledge
- 9) Ability to present quantitative information (specific data, tables, graphs, in reports and similar)
- 10) Using data ethically
- 11) Ability to apply results to learning, decision making or problem-solving
- 12) Ability to plan, organize and self-assess throughout the process

2.5.3 The Third Phase (A Conceptualization Phase)

In the third phase, a significant number of discussions on DL have been constantly emerging, including studies on data citizenship and new required capabilities for DL. With the rapid development of open data movement calls for all citizens to have some level of DL (Twidale et al., 2013; Wolff, Gooch, et al., 2016; Wolff, Montaner, & Kortuem, 2016). Facing the idea of “global data literacy”, Gray et al. (2018) recommended an expansion of the concept of DL:

“to include not just competencies in reading and working with datasets but also the ability to account for, intervene around and participate in the wider socio-technical infrastructures

through which data is created, stored and analyzed.” They call it “data infrastructure literacy.”

In terms of this idea, they suggested considering data infrastructure literacy as a site for ongoing public engagement and experimentation around infrastructures of datafication. This concept demands higher requirements for DL.

In addition, reflections on data citizenship argued that data inequality exists in the gaps of people with different levels of data literacy (Carmi et al., 2020; D’Ignazio, 2017). The inequality can take two forms: 1) the inequality between those who are proficient in data-domain knowledge, e.g., storage and collection, and those who are not (Andrejevic, 2014), and 2) the inequality between those who are proficient in the technical skills required to effectively work with data and those who are not (D’Ignazio, 2017). Also, Carmi et al. (2020) proposed the idea of rethinking data literacy in the age of disinformation, misinformation, and malinformation. They argued three paramount data literacy presented by them from previous studies, including citizens’ critical understanding of data, citizens’ everyday engagement with data, and citizen’ proactive engagement with data and their networks of literacy.

During this phase, extant studies or reports have used geospatial data to conduct research analysis, design innovations, or tell stories, such as investigating economic and environmental relations (Pászto & Zimmermannová, 2019), designing smart cities (Audu et al., 2019; Walravens et al., 2014), or simply presenting parks’ distribution in a city. Given these usages, geospatial data literacy was brought into the data literacy world. Juergens (2020) claimed that geospatial data are a specific type of data and influence most of our daily decisions, but few people are proficient in processing geospatial data. This study also points out the capabilities of geospatial data literacy, namely raster cell size and map projection. Kremer et al. (2022) raised the significance of improving data literacy in spatial disciplines and identified critical digital tools for spatial analysis, such as QGIS and Geopandas.

The world is growing so fast that technologies have been consistently changing the world, so the understanding of DL and the required DL capabilities will inevitably evolve. Considering the evolvability of the definition of DL, constant consideration and discussion of this definition will be advantageous to ensure all aspects of this literacy are addressed. As the

evolution of DL was presented in this section, in the initiative phase, scholars viewed DL only from the perspective of data consumers; the requirements were more connected with SL and IL and were not very specific. Moving into the next phase, since 2010, the concepts of DL began to incorporate the perspectives of data consumers and data producers; plus, with the rapid development of technologies, the requirements of a data literate have greatly expanded. Nowadays, the open data movement demands even higher requirements for DL, and especially for the need that all people should have some level of DL. It is a big challenge to design DL programs and evaluate them. Given the new and inevitable demand for DL, there is a need to keep investigating DL, primarily focusing on open data literacy, which can be used by anyone alike, including researchers, data scientists, and citizens.

3.0 Research Questions, Research Design and Conceptual Model

This chapter presents the research questions of this dissertation (Section 3.1), and illustrates the corresponding research design that is devised to answer the research questions (Section 3.2). Section 3.3 describes my initial conceptual model for H-OGD-I which was built upon an existing HDI framework and sensemaking model. This initial conceptual model was then used to develop research instruments and create the content analysis schema.

3.1 Research Questions

To achieve the objectives, I propose to answer three sets of research questions (RQ). RQ1 addresses Objectives 1 and 2, which are to explore typical user behaviors and develop an evidence-based H-OGD-I model. RQ2 undertakes Objective 3, which is to reveal contextualized user challenges and disclose the linkage between those challenges and OGD literacy. RQ3 achieves Objective 4, which is to investigate the OGD literacy capabilities that enable users to use OGD and develop a corresponding taxonomy.

RQ1: What are the typical user behaviors when interacting with OGD?

RQ1.1. What types of tasks are performed when users interact with OGD?

RQ1.2. What are the users' accessing behavior patterns when interacting with OGD and its portal websites, e.g., accessing channels and seeking behavior patterns (browse and keyword search)?

RQ1.3. What are the typical user behaviors in each interacting stage?

RQ2: What are the challenges users face when they interact with OGD?

RQ3: What are the fundamental OGD literacy capabilities that enable users to use OGD?

3.2 Overview of Research Design

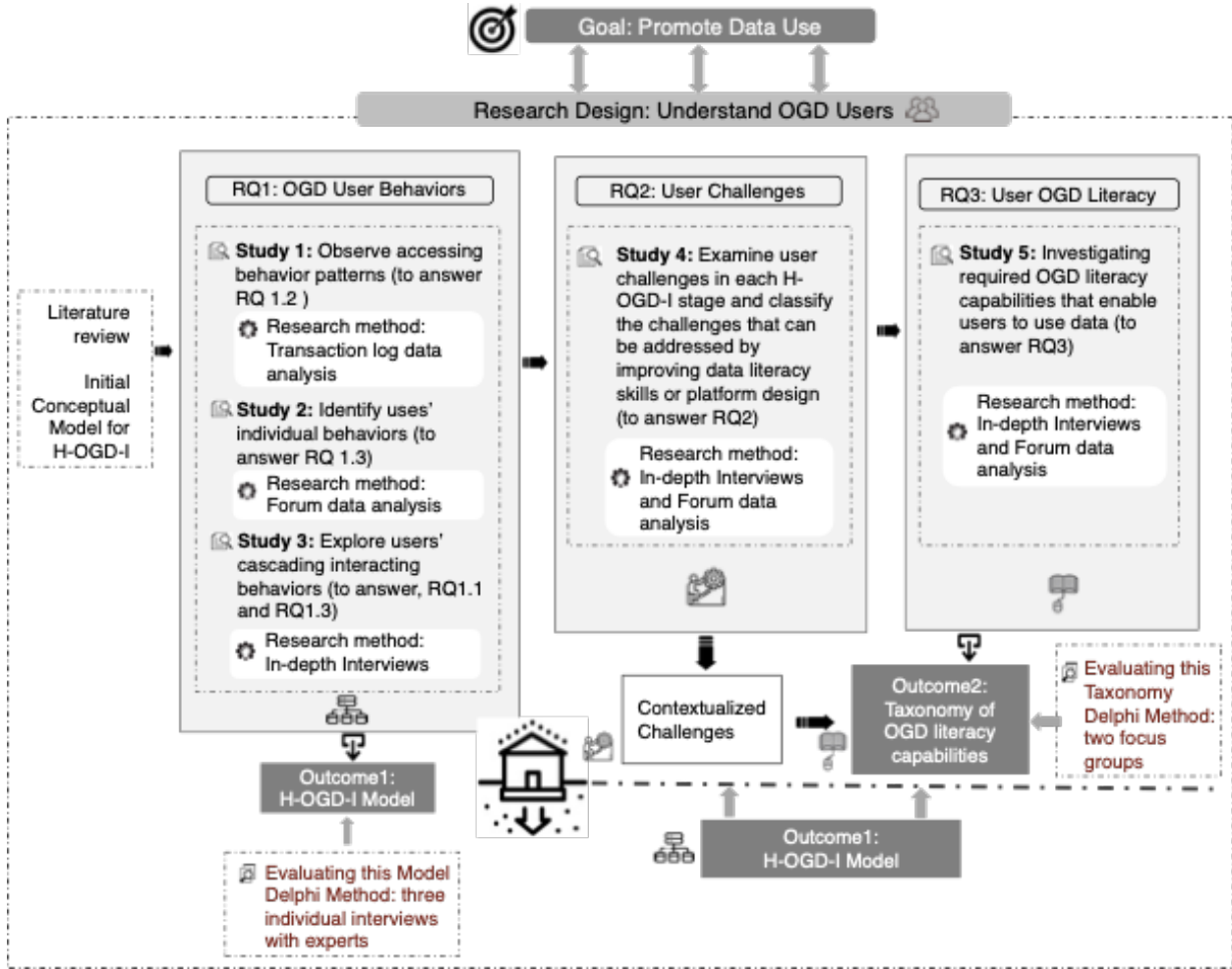


Figure 10: Overview of Research Design

This dissertation aims to promote data use from a user-centered angle. I proposed three RQs concentrating on understanding OGD users, including user behaviors, user challenges, and user literacy. In addition, the final two main outcomes from this dissertation contain a theoretical achievement - a model for H-OGD-I, and a practical achievement - a taxonomy for OGD literacy capabilities. To achieve the goal and outcomes, this dissertation adopted a mixed methods design which refers to the equal use of quantitative and qualitative methods during the same time frame in order to study the same research questions (Creswell, 2014). The reason for collecting both quantitative and qualitative data is to merge and validate

the results of the two forms of data to bring greater insight into the problem than would be obtained by interpreting either type of data separately. Figure 10 demonstrates the overview of the research design and highlights essential studies and the corresponding study sequences.

To explore how humans interact with OGD, I first proposed an initial conceptual model for H-OGD-I based on the literature review, and this initial model was applied to the following three empirical studies (Study 1 & 2 & 3) to guide the designs of instruments and data analysis methods. The findings from these three studies contributed to the final adjustments of the H-OGD-I model. Study 1 used a transaction log analysis approach to answer RQ1.2 related to the user accessing behavior patterns, including the channels users adopted to visit OGD portal websites and behavioral preference of browsing or keyword searching. Study 2 adopted the forum posts (content) analysis (conducting a content analysis on an online discussion forum posts) method to answer RQ1.3 regarding users' individual OGD interacting behaviors. Study 3 utilized the interview method to answer RQ1.1 about the types of tasks users perform when using OGD, and RQ 1.3 concerning users' cascading OGD interacting behaviors. Studies 1 & 2 were conducted concurrently but separately to help me best understand the research problem (Creswell, 2014). After obtaining results from Studies 1 & 2, an in-depth interview method was adopted (Study 3), which was enlightened by the explanatory sequential design, allowing for the collection of statistical results from the quantitative study; following up with individual interviews assists in explaining those results in more depth.

To identify the contextualized users' challenges when they interact with OGD and point out the challenges that can be addressed by improving data literacy skills or platform design (RQ2), two qualitative approaches were utilized (Study 4), including forum posts data analysis and interview studies with OGD users.

Based on the findings from Study 4, Study 5 was performed by using the research approaches of forum posts analysis and interview studies to discover the fundamental OGD literacy capabilities that allow users to effectively use data (RQ3),

Finally, the Delphi method was applied to evaluate the dissertation findings. Three experts from OGD domain and HDI field were invited to evaluate the proposed H-OGD-I model. Also, two focus groups were conducted to evaluate the taxonomy of OGD capabilities.

The experts in the focus groups are in the fields of OGD literacy and data science.

3.3 Conceptual Model for H-OGD-I

The literature review (Chapter 2) reveals that the OGD project is still in its early stage, and the research field of HDI is just emerging in OGD practice. However, as the emphasis of OGD projects has shifted from a publisher-centered paradigm to a user-centered paradigm, and a massive number of datasets are available online, the research field of HDI will be a prosperous area in OGD user behavioral studies. Also, it will undoubtedly play an essential role in contributing to OGD use, thereby generating great social, political and economic value. Nonetheless, currently, there is a lack of research on HDI topics and even fewer studies that investigate applying HDI to OGD practice.

In addition, even though HII has long formed a principal focus of research efforts and has developed a series of mature models, the theories or models of HII cannot be directly applied to the HDI field for two reasons. The first reason is that *data* is different from *information*. *Data* are raw materials, e.g., numbers, characters, and symbols, which do not contain meanings, whereas “information is data rendered meaningful via analysis and structuring” (Ford, 2015), which embraces meanings. Data can be converted to information by combining and analyzing various data. Consequently, raw data can be challenging to understand and therefore needs an additional process to interpret the data. The second reason HII models cannot be directly applied to the HDI field is the different connotations between the terms *interaction* and *behavior*. In LIS, to study behavior is to investigate observable and identifiable information-related activities and to have no claims on understanding internal, unobservable processes and perceptions (Fidel, 2012; Wilson, 1994). However, *interaction* focuses not only on the phenomena but also the dynamic nature of the relationships between people and information that is a connotation missing from the term *behavior* (Fidel, 2012). Also, most of the theories or models in HII are concerning HIB, and most HIB models are information seeking behavior models without addressing the complete interaction processes; thereby, I determined to leverage sensemaking models to explain the process of interpreting

data and fill this missing connotation of *behavior*.

To formulate a conceptual model of HDI within the context of OGD, I integrated L. Koesten et al. (2017)’s HDI framework and Pirolli & Card (2005)’s sensemaking model for two pivotal considerations. First, because Koesten et al.’s framework focuses on users’ online data-seeking behavior, and they studied it in the context of OGD, we considered their framework a good start to help develop a comprehensive model. Second, because data are raw materials, e.g., numbers, characters, and symbols, which do not contain meanings (Ford, 2015), raw data can be challenging to make sense of, and therefore needs an additional process to interpret the data. To address the additional process, we decided to leverage a sensemaking model to explain the process of interpreting data. As Pirolli and Card’s sense-making model takes both sensemaking behaviors and information-seeking behaviors into account, we also adopted their sensemaking model. These two models have some overlaps and complement each other; we expect that integrating them can describe the entire process of interacting with data, as shown in Figure 11.

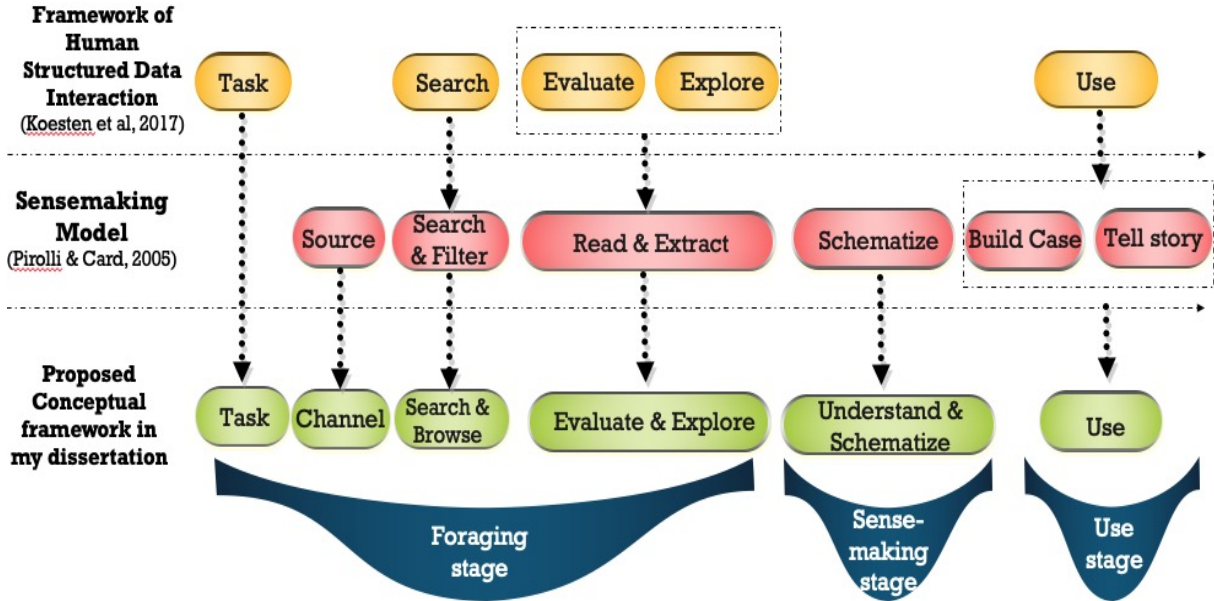


Figure 11: The Proposed Conceptual Model of HDI

This initial configured model depicts the whole process of human interaction with OGD, which is comprised of three main stages: foraging stage, sensemaking stage, and using stage.

Specifically, the foraging stage involves the components of the task, channel, search & browse, and evaluate & explore. The sensemaking stage includes the processes of understanding and schematizing. The using stage refers to using the data for the project. Each component is described as follows.

Task refers to the motivation of initiating a data interaction. In this dissertation, a task is defined as a series of “activities to be performed in order to accomplish a goal” (Hansen, 1999).

Channel indicates how people found the OGD portal’s website. When people try to find a dataset, they may choose to use a search engine to search, e.g., Google; or directly search data in a known data portal.

Search & Browse denotes the search strategies when people begin to look for data. In general, there are two main search strategies: *keyword searching* and *browsing*. In this dissertation study, *keyword searching* is considered when individuals are able to form queries, enter a query on the OGD portals and perform either breadth-first move or depth-first move. They may keep refining their queries to achieve the final goals. *Browsing*, on the other hand, is defined as a set of intentional seeking activities on online systems (the goal could be clear or vague), where these seeking activities include the use of the navigations or filters provided by the platforms rather than forming their own keywords/queries to search the needed information. The main difference between *keyword searching* and *browsing* is that keyword searching requires users to know what the appropriate keywords are to do the search, while the users who choose to browse to obtain the needed data rely more on the interface design, including the navigations or the filter functions. If the searching behaviors trigger the navigation or filters, it will be considered as a combination of browsing and searching.

The understanding of *Evaluate & Explore* stage references the description of L. Koesten et al. (2017)’s framework. *Evaluation* is described as the first look when users evaluate the relevance of the identified dataset with their data needs. The actions could involve examining the title and description of the dataset. Then, with further *exploration*, users begin to build a notion of quality and trust of the data by reviewing related information and checking data quality, e.g., examining metadata. Ultimately, users need to decide whether or not go deeper into this dataset. The two stages are different because *exploring* happens after *evaluating*;

it is a second look, but closely connected because when users consider that the identified data is relevant to the data needs in the stage of evaluation, they will immediately enter the exploration stage.

The stage - *Understand and Schematize* - signifies the process of sensemaking. Based on the definition of OGD, users are facing the raw structured data, which may not be meaningful if they only see the numbers. Therefore, before using it, they have to understand it first. *Understand* signifies the moment when users start to perceive the intended meaning of the data. Typical behaviors include closely reading the codebooks or data dictionaries or even looking for related concepts from other websites. The interpretation of *Schematize* is implied by Pirolli & Card (2005)'s sensemaking model, which indicates users try to organize the data with relevant information or build connections between the data to schematize a simple schematic form of understanding.

The last stage is when people began to *use* the identified data to carry out their tasks. The possible activities include data cleaning, data merging, data analysis, and finally achieving the goal of the task.

This initial conceptual model was first applied to three subsequent studies with the intent to further devise the research instruments and build the content analysis schema. The findings from these three studies contributed to the final adjustments of the initial model to form the H-OGD-I model used to understand user behaviors when interacting with OGD.

4.0 Dissertation Studies

4.1 Research Sites and Research Population

This dissertation focuses on local-level rather than federal- or state-level OGD portals and their users. As Conradie & Choenni (2014) proposed, OGD is mainly collected at the local level; thus, supporting local-level OGD is supporting the success of OGD as a whole. They also argued that to best understand the concept of OGD, investigating local-level OGD is a better start (Conradie & Choenni, 2014; Janssen et al., 2012). A local-level OGD portal is critical because of its closer connection with local organizations, neighborhoods, and communities, which could directly impact a citizen’s daily life and neighborhood. However, many extant works emerge from a national perspective, and few studies focus on investigating local-level OGD (Conradie & Choenni, 2014; Janssen et al., 2012). Therefore, this dissertation tends to fill in this gap.

This dissertation project studies the users from three local-level OGD data portals, two of them located in two major cities in Pennsylvania: the portal from the city of Philadelphia (OpenDataPhilly) and Western Pennsylvania Regional Data Center (WPRDC), the other one is located in the city of Boston (Analyze Boston). All the three portals are built on CKAN platform. OpenDataPhilly, WPRDC and Analyze Boston not only provide access to free datasets but also offer various tools and visualizations to make the data easier to find and use.

OpenDataPhilly is “a catalog of open data in the Philadelphia region. In addition to being the official open data repository for the [c]ity, it includes datasets from many organizations in the region.”¹ To date, OpenDataPhilly contains a total of 386 datasets across topics including the environment, health, and real estate, to name a few. WPRDC was established in 2015 and is managed by the University of Pittsburgh Center for Urban and Social Research, in partnership with Allegheny County and the City of Pittsburgh. According to its website, WPRDC “maintains Allegheny County and the City of Pittsburgh’s open data

¹OpenDataPhilly, <https://www.opendataphilly.org/about>

portal and provides a number of services to data publishers and users. The Data Center also hosts datasets from these and other public sector agencies, academic institutions, and non-profit organizations.”² To date, WPRDC reposit 331 datasets in total, that cover the topics of environment, health, and transportation, to name a few. Analyze Boston is the City of Boston’s open data hub. According to its website, Analyze Boston is “working to make this the default technology platform to support the publication of the City’s public information, in the form of data, and to make this information easy to find, access, and use by a broad audience.”³ To date, Analyze Boston stores 233 datasets containing similar topics to the other two OGD portals, namely, environment, public health, and public safety.

In this dissertation, the targeted users are the non-expert end users who have experience interacting with OGD. An end user refers to an individual who uses the OGD directly rather than consuming the effects of OGD application, e.g., by using transportation apps.

4.2 Data Collection for the Three Datasets

This dissertation contains five sub-studies, as shown in the Research Design section (3.2). These five sub-studies utilized three datasets collected from the three local OGD portals via different methods. As shown in figure12, transaction log data were collected from all three portals, OpenDataPhilly, WPRDC, and Analyze Boston; Online forum posts data were collected from an online discussion group managed by OpenDataPhilly; most interview data were collected from the users of WPRDC and OpenDataPhilly. Even though the datasets are the same, sub-studies analyzed different parts of the datasets to draw conclusions. Figure 12 illustrates the overview of data collection methods and their application to sub-studies. The following sections depict the data collection methods for each dataset.

²WPRDC, <http://www.wprdc.org/about/>

³Analyze Boston, <https://data.boston.gov/>

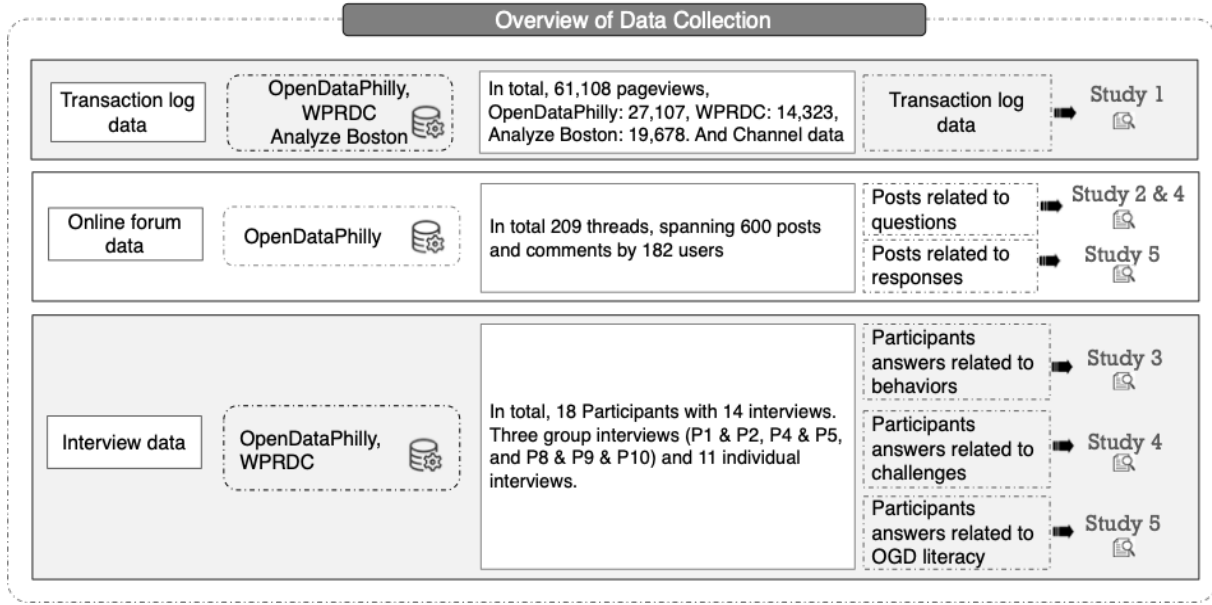


Figure 12: The Overview of Data Collection

4.2.1 Transaction Log Data Collection

The transaction log data were provided by the three research sites (OpenDataPhilly, WPRDC and Analyze Boston). All three portals use Google Analytics to generate and store their transaction log data, and the data within the three portals are automatically logged. The data were collected from the same time period - January 1, 2018 to June 30, 2018. This study focused on two categories of data: pageview data and channel data. Based on Google Analytics, pageview data is defined as an instance of a page being loaded (or reloaded) in a browser.⁴ Channel data refers to the categories that indicate how people found the website, such as organic search or referral. In total, the transaction logs from the three portals contains 61,108 pageviews, where OpenDataPhilly has 27,107 and WPRDC has 14,323, ABoston has 19,678. The channel data is recorded in an independent table, which can be directly to compute the commonly used visit channels.

⁴Google Analytics, <https://support.google.com/analytics/answer/6086080?hl=en>

4.2.2 Online Forum Posts Data Collection

The online forum posts data were manually collected from the Discussion Group for Open Data and Government Transparency in Philadelphia to reveal user behaviors, challenges and OGD literacy capabilities when interacting with OGD. This forum is a public question-oriented online discussion group managed by OpenDataPhilly, which provides a platform for communications between OGD users and the organizers of this portal. This platform allows the public to ask questions about the location of specific data, data interpretation, as well as to bring up challenges they encountered. Figure 13 shows a screenshot of the homepage, and figure 14 presents one example of the posts in this group. The data was collected from January 2019 to February 2022, containing 209 threads, spanning 600 posts and comments by 182 users. Specifically, the information of the post content, post IDs, post times, as well as the corresponding responses and timestamps was collected (see example in figure 15). In providing further privacy protection, I used meaningless strings to create user IDs, whereby the user ID is used for sole purpose of separating the posts by different users. This work was approved by the Human Research Protection Office of the University of Pittsburgh (PittPRO).

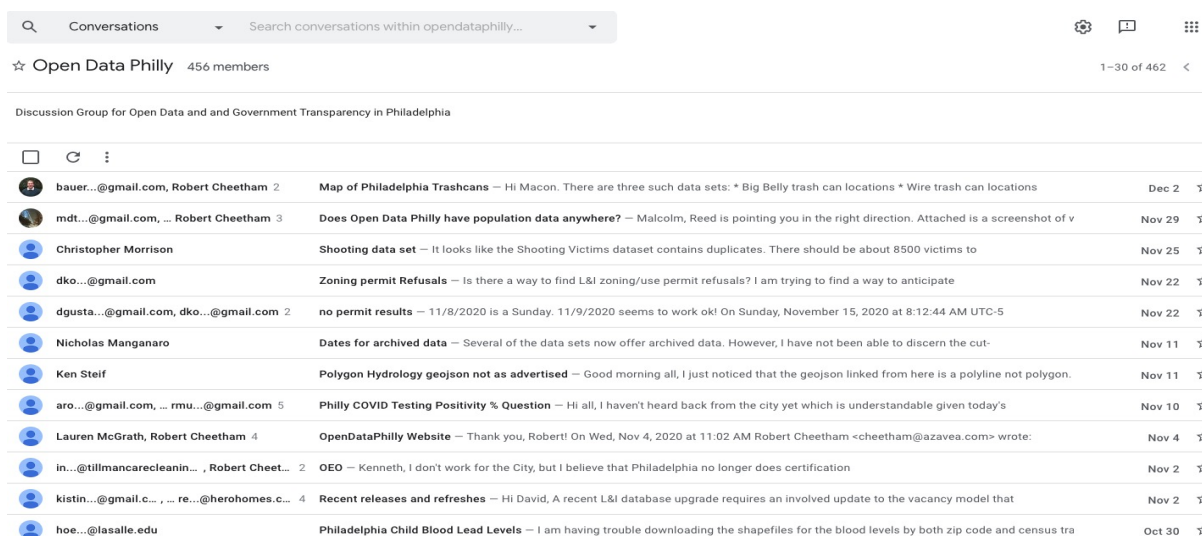


Figure 13: Homepage of Discussion Group for Open Data and Government Transparency in Philadelphia

Map of Philadelphia Trashcans 8 views

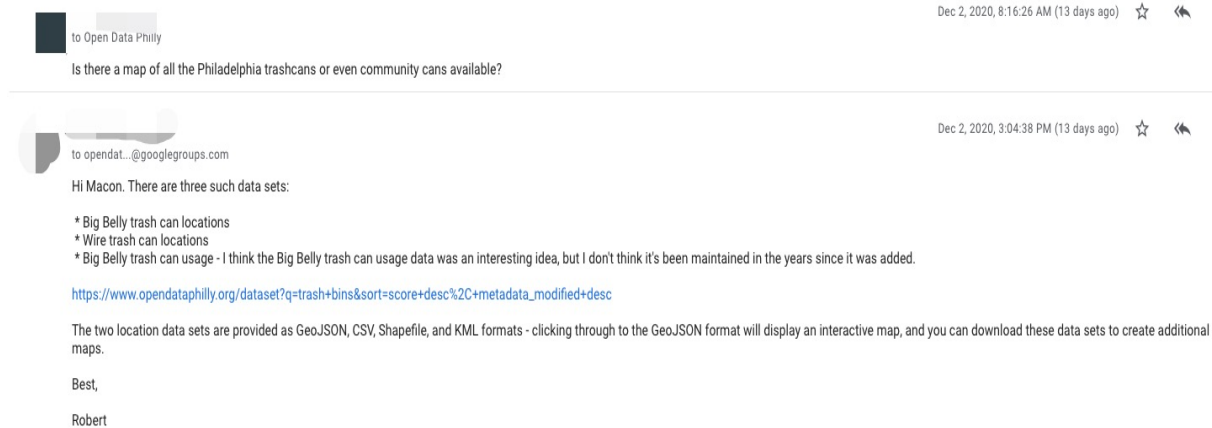


Figure 14: An Example of Posts

Post 1

user_id: u00001

article_id: a00001

action: post

response_id: null

organization: The Bicycle Coalition of Greater Philadelphia

time: 1/3/2020

title: Traffic Fatalities not updated since November 24th

content: Just a heads up that the data on the Traffic Fatalities is not refreshing.

user_id: u00002

article_id: a00002

action: response

response_id: a00001

organization: OpenDataPhilly

time: 1/10/2020

title: null

content: Thank you for bringing this to our attention. The data should be up-to-date now. Please note that the links on OpenDataPhilly are new so you might need to update any connections you've made.

Figure 15: An Example of Group Data Collection

4.2.3 Interview Data Collection

Interview technique. To explore a wide range of insights about users' experience of interacting with OGD, the critical incident technique (CIT) was adopted as the basis for designing a semi-structured interview instrument (Flanagan, 1954). Under this interview approach, each participant was asked to recall one incident in which they interacted with OGD. Recollection of participants' own word choice experiences allowed us to understand real tasks, restore natural processes, and have a deeper understanding of corresponding user behaviors and challenges and the needed OGD literacy while interacting with OGD. The initial H-OGD-I model was applied to design the interview questions. Finally, the unrevealed patterns/questions/challenges/ found in Study 1 and 2 were utilized to develop a list of prompts to help the participants remember specific incidents from their memories and capture the details in the data interaction process.

Interview data collection. In the application of CIT, the first and foremost requirement to qualify participants is that they must have recent experience with OGD. Then the snowball sampling technique was used to recruit more participants (14 participants). Additionally, the director of WPRDC helped us disseminate the interview invitation letter to their users (2 participants). An invitation letter was also posted to the Discussion Group for Open Data and Government Transparency in Philadelphia (2 participants). In total, 18 participants were interviewed, and each interview lasted around 40 to 60 minutes. Notably, I conducted three group interviews (P1 & P2, P4 & P5, and P8 & P9 & P10) and 11 individual interviews because the participants in each group collaborated to complete the same project, and they were assigned different responsibilities for the tasks. Therefore, interviewing them in a group can obtain a more comprehensive description of the process. Even though it was a group interview, I asked all the participants to engage in every question.

Samples. An interview question asked the participants how many times they have used OGD. Among the 18 participants, the distribution is 1-5 times (8), 6-15 times (5), 15 – 100 times (3), and more than 100 times (2). This distribution of the participants allows us to observe users' behaviors covering different levels of familiarity with the OGD. Furthermore, during the process of interviewing participants, repetitive answers were continuously coming

out, no matter a new user or a user with a rich experience of interacting with OGD. This phenomenon further confirmed that the interview sample size is enough to qualitatively discuss user OGD behaviors.

Instrument design. The interview questions were structured into four main sections. The first section consists of questions concerning participants' background information: for example, demographic information, such as participants' current primary professional positions, disciplines, and the frequency of using OGD. The questions in the second section are designed based on the initial H-OGD-I model, covering the entire process of the data interactions. A set of questions regarding the challenges encountered in each interacting stage were followed up in the third section of questions. The fourth section's questions are related to the required OGD literacy that assisted participants in overcoming the challenges. This study design was approved by the Human Research Protection Office of the University of Pittsburgh (PittPRO).

4.3 Study 1: Observing Users OGD Accessing Behavior Patterns

Regarding transaction log data collection, please see section 4.2.1.

4.3.1 Data Processing and Analysis

This study examines users' data accessing behaviors patterns, including what channels users use to visit portals' websites, their browsing and keyword search preferences when seeking for data, and the preferred browsing entries to datasets, such as topics and formats.

Channel data and pageview data is used to identify what channels user adopted to visit OGD portals' website.

Due to the transaction logs structure, to observe users' accessing behaviors, the URLs of the pageview events need to be dissected. For example, event URL `/dataset?organization=city-of-philadelphia&page=1` shows that the user was browsing by clicking the link labeled City of Philadelphia under the category of Organization, or event `/dataset? q=census+tract,`

shows that the user was searching by using the query of “census tract.” Table 1 shows the URL examples of browsing behaviors.

Table 1: Browsing by Different Entries

Browsing by	Key part of URLs	Examples
Organizations	/dataset?organization or /organization	/dataset?organization=city-of-philadelphia&page=1
Topics	/dataset?groups= or /groups	/dataset?groups=environment&tags=social+determinants+of+health
Tags	/dataset?tags=	/dataset?tags=abandonment
Format	/dataset?res_format=	/dataset?res_format=XLS
License	dataset?license_id=	/dataset?license_id=odc-by
Data request	/datarequest	datarequest/e19abf49-aaca-4c90-948d-8a088467532a

Sometimes, a user might combine search and browsing in one action. For example, in URL: /dataset?q=1015+south+ave+&tags=assessment&organization=Allegheny, the user searched with the query “1015 south ave” under the tags assessment and the organization of Allegheny.

After identifying the URLs, a process of data extraction and computation was performed to obtain the frequency data of users accessing behaviors, including browsing, searching, and the combination, respectively. Also, the frequency of browsed entries was counted in order to know the preferences of browsing categories. For example, how many users prefer to access data through a category of Organizations or Topics. Since we focused on the user’s accessing behavior, we only computed the frequency of unique URLs for each user.

Our data analysis needed to differentiate successful access from a failed one. Since the goal of users’ OGD access is to find the right data to use, I assume in this study that an event action download signifies that the user has found the data he or she wants. Consequently, any access session (either via searching with a query, browsing within the multi-facet interface or the combination of the two) is successful when there is a download action, whereas a failed one does not. I acknowledge that this approach could wrongly classify some access sessions. For example, a user may download a dataset to later realize that it is not what he

or she needs, or a user might obtain the information he or she needs without downloading the dataset. However, due to the lack of more accurate measurement than download events in the log, such an assumption had to be made. More importantly, this assumption probably works the majority of the time, so it would be adequate for the purpose of studying the common patterns of large numbers of users' OGD accessing behaviors.

4.4 Study 2: Identify Users' Individual OGD Behaviors

Regarding online forum posts data collection, please see section 4.2.2

4.4.1 Data Analysis

After examining all the 600 posts, 362 posts unrelated to user behaviors were excluded. For example, the post contains an announcement about releasing a new dataset or answering a question. Finally, we identified 238 posts that directly manifest user behaviors. The 238 posts were further annotated based on the themes of OGD users' interactive behaviors.

This content analysis was conducted by me and other two researchers through an iterative process, which contained three main steps. Firstly, I developed an initial coding schema based on the initial H-OGD-I model. Secondly, a first-round random data coding was conducted to adjust the schema, and during this round, the three researchers discussed the coding schema together to finalize it. Thirdly, based on the finalized coding schema, another researcher and I went through four rounds of the independent annotation process. For the first three rounds, we annotated the same group of posts, discussed our results, and obtained a unanimous agreement with our coding results. Finally, I annotated 238 posts (100%), the other researcher coded 159 posts (67%), and our Kappa score reached 0.93, which is a satisfactory value (Viera et al., 2005).

4.4.2 Coding Schema

This coding schema was developed based on the initial conceptual H-OGD-I model and then refined according to the analysis of forum posts analysis and interview data. After several rounds of annotation and discussion, the coding schema was formulated to contain four behavioral stages: data forage, data sensemaking, data use, and data share. Data forage covers the processes of finding, evaluating, scrutinizing, and acquiring data. Data sensemaking contains the processes of understanding and schematizing data. Data use comprises the processes of preparing data, analyzing data, representing results, interpreting, communicating with data, and making data-driven decisions. Data share, which was not included in the initial conceptual model, was discovered from the forum data analysis, and added to the coding schema. The following elaborates on the coding schema and provides the corresponding examples.

* Stage of Data Forage

Find (code: FIND) denotes the process when users look for data, including identifying data sources and applying data-seeking strategies to find data.

Evaluate (code: EVA) denotes the first look when users examine the topic relevance of the identified dataset with their data needs. The actions could involve reading the title and description of the dataset.

Scrutinize (code: SCRUI) denotes the second look when users rigorously scrutinize the utility of the identified dataset, including examining the file formats, data types, and data accuracy; assessing the variables that can combine multiple datasets.

The following are the sub-codes for Scrutinize behavior.

Sub-code: SCRUI-Time. Examine the specific time range covered by the datasets. *Example: “I used to use this dataset for real estate analysis and research. I see the most recent sale dates end in 2020. From the activity log, looks like the last update was 10-11 months ago. Does anyone know how this dataset can be updated?”*

Sub-code: SCRUI-Content. Examine the specific content of a dataset, namely the scope of a dataset. *Example: “ Does anyone know if there is a possibility that older crime incidents are now being added to the dataset?”*

Sub-code: SCR-Accuracy. Examine the accuracy of a dataset, namely sparse and duplicated data. *Example: “I have been eyeing the tracking of COVID death data per date and noticed that there is increasingly sparser data the last few weeks and repeat values.”*

Sub-code: SCR-Missing. Examine the data fields and values. The dataset topic is relevant, but the needed data fields are missing. Or, the data fields exist, but the corresponding values are missing without any explanation. *Example: “When I opened the 2016 csv file, I couldn’t see any data other than object ID. I am particularly interested in tree species.”*

Acquire (code: AQU) denotes the process when users obtain the datasets. The behaviors may involve clicking the provided link to download the datasets directly and querying data through APIs.

The following are the sub-codes for Acquire behavior.

Sub-code: AQU-URL. The OGD users acquire data by using the URL provided by portals’ website. *Example: “Looks like none of the data download links are working. I’m getting ‘500 Internal Server Error’ when I try to download. Can you help?”*

Sub-code: AQU-API. The OGD users acquire data by using the API provided by portals’ website. *Example: “Could you explain how to use the API to access the databases? There’s another one in opendataphilly that gets stuck downloading the CSV.”*

Sub-code: AQU-Manu. The OGD users acquire data by manually collecting the data from portals’ website. *Example: “I hadn’t thought about a chron job (and honestly with my skills, it would take me a ton of trial and error to get it done), so I have just been manually cutting and pasting.”*

Sub-code: AQU-WAT. The poster tried to acquire data by using some web applications or tools, such as Tableau. *Example: “I then tried web application and sorted on a single census tract. Was not able to download census tract level OPA data. Any thoughts on next step?”*

* Stage of Data Sensemaking

Understand (code: UND) denotes the process when users start to perceive the intended meaning of the data, such as understanding the meaning of the column headings (field names) and the values of the data. Typical behaviors include closely reading the data

dictionaries and/or ‘read me’ files or looking for related concepts from other websites. *Example: “Does anyone know what the shape_area and shape_length data represents in the Philly zipcode dataset?”*

Schematize (code: SCHE) denotes the process when users try to establish the logical relationships between the data. For example, organizing the data with relevant information or building connections between the data to schematize a simple schematic form of understanding. Typical behaviors include sorting or filtering to find connections between data within one dataset or multiple datasets. *Example: “I see polygons- labeled Loss, Gain, or No Change- but I’m wondering how those characteristics affect the area of the polygons.”*

*** Stage of Data Use**

Prepare (code: PREP) refers to the process of making the data ready to be further analyzed, visualized, or directly used (e.g., building a database). The specific activities could involve cleaning, extracting, importing, transforming, and merging. *Example: “Anyone have a quick instruction on importing or linking to maps showing police districts for Philadelphia? I have tried different approaches, but I haven’t gotten far.”*

Analyze (code: ANALYZE) denotes the process of utilizing various methods (mainly quantitative) to analyze data to discover useful information, identify patterns, draw conclusions, and/or support decision-making. *Example: “the 10/29 numbers are 4111 Negative and 447 Positive which by my math would be 4558 Total with a Positivity % of 9.81% (double-checked by percentagecalculator.net) but the “Positivity%” purple line graph for 10/29 lists 7.7%.”*

Represent results (code: REPRES) depicts the process of representing the results through different methods. *Example: “I’m linking my spreadsheet this time...”*

Interpret (code: INTERP) indicates the process of exploring and translating the analysis results through combining the comprehension of domain knowledge and the data analysis results to finally draw conclusions or make decisions. *Example: “In terms of the OpenData field, isn’t this an unusually large overnight data change indicating some sort of previous issue with an important dataset?”*

Communicate (code: COMM) manifests the process of using appropriate language and manners to communicate the data with targeted audiences, such as telling an evidence-

based data story. *Example: “Any review or feedback or double-checking appreciated especially since I have no understanding of carto etc.”*

Making decisions (code: MD) describes the process of making a choice or deciding to take action based on understanding of the identified data or the interpretations of the analysis results. *Example: “...So I resorted to using streetmartphl to figure out when to put my trash out. Looking at adjacent zones for 2-3 days, you can start to see patterns and it becomes fairly clear which day the garbage truck is coming to your block.”*

*** Stage of Data Share**

Share data (code: SHARE) denotes the process when users share the data collected from various data sources through different platforms and methods, e.g., GitHub, Google sheets, and/or emails. This behavior stresses that data sharing happens between users rather than data portals or organizations. *Example: “I started additional cron jobs to cover all the datasets. I’ll start figuring out how to publish them to Github in an organized manner.”*

This coding schema was also applied to the interview data (Study 3) analysis.

4.5 Study 3: Explore Users’ Cascading OGD interacting behaviors

Regarding interview data collection, please see section 4.2.3

4.5.1 Data Analysis

In Study 3, the 18 participants contain 10 master’s students (7 in Library and Information Science, 1 in Information Science, 1 in Urban Studies, 1 in Humanities), 2 Ph.D. students (1 in Library and Information Science, and 1 in Information Science), 1 data analyst, 1 faculty member, 1 researcher, 1 principal engineer, 1 urban planner, and 1 software engineer. The age groups are comprised of 18-24 (3), 25-34 (9), 35-44 (4), and 45-54 (2). As mentioned in section 4.2.3, the participants cover the different levels of familiarity with the OGD, from knowing a little (1-5 instances of use) to being very familiar (more than 100 instances of use). This study uses P plus a number to represent the participants, namely P1.

All interviews were recorded and transcribed. Two researchers and I conducted an inductive and qualitative thematic analysis using the same coding schema (4.4.2) as the Study 2. I coded and analyzed all the interview data and discussed the results with the other two researchers to achieve a unanimous agreement.

4.6 Study 4: Examine User Challenges in each H-OGD-I stage

This study aims to contextualize user challenges when they interacting with OGD by analyzing both forum posts and interview data, thereby, regarding data collection methods, please see the section of 4.2.2 and 4.2.3.

4.6.1 Data Analysis

Based on forum posts (content) analysis (Study 2), user challenges were able to be manifested by the posts associated with asking questions and seeking help. We found the 238 posts identified for user behaviors also indicate users' challenges. For example, "*Can anyone point me to datasets the City may have available with regards to street parking? In particular I am looking for: 1. Location of all fire hydrants 2. Location of all non-parking zones (a superset of 1) 3. The location and designation of all parking signs.*" This post presents the behavior of *find*, also indicates that users encounter challenges in finding data, e.g., do not know where to find the specific data. Therefore, we used this same group of posts to further analyze OGD user challenges.

Another dataset we used to identify OGD user challenges is the interview data. As stated in the section on interview instrument design (section 4.2.3), a set of questions regarding the challenges encountered in each interacting stage was designed in the third part of the questions in the instrument. After participants described a process of behaviors, a follow-up question concerning what challenges they faced in this stage was asked.

Based on the new H-OGD-I model (section 5.1.5), and referencing the coding schema for OGD user behaviors, I developed a new coding schema for OGD users challenges, shown

in Table 2 . The content analysis process was similar to Study 2 (forum post analysis). A first-round random data coding was conducted to adjust the initial coding schema. In the meantime, another researcher, my adviser, and I discussed the schema together to refine the categories and the corresponding explanations. At last, another researcher and I annotated the posts utilizing the finalized version codebook. I coded all the posts, and another researcher randomly coded 25% posts. Cohen’s Kappa agreement score is 0.88, which indicates a substantial agreement(Viera et al., 2005) between the two coders. In addition, I also coded and analyzed all the interview data and discussed the results with the other researcher and my adviser to achieve a unanimous agreement.

4.6.2 Coding Schema

Table 2: The Coding Schema for OGD User Challenges

Behaviors	Description
Find code: FIND	The post presents the challenge when users cannot find the needed data. E.g., <i>“I am looking to see if there is a data set available which plots out all of the benches in the city. Does this exist and if not what next steps should I take?”</i>
Scrutinize code: SCRUI	The post presents the challenge when users critically scrutinized if the identified dataset is usable for their project. E.g., <i>“It looks like the Shooting Victims dataset contains duplicates. There should be about 8,500 victims to date, but there are 17,000.”</i>
Acquire code: ACQ	The post presents the challenge when users acquired the datasets. E.g., <i>“Could you explain how to use the API to access the databases? There’s another one in opendataphilly that gets stuck downloading the CSV.”</i>
Understand code: UNS	The post presents the challenge when users cannot understand data. E.g., <i>“anyone know what these values represent? they’re integers -1 through 9, there’s no description on the metadata”</i>
Schematize code: SCHE	The post presents the challenge when the user cannot understand the logic of the data to establish the relationships between data. E.g., <i>“Any insight on why homestead exemptions might vary so significantly?”</i>
Use code: USE	The post presents the challenge when users try to clean, import, analyze data or other types of using data. E.g., <i>“Anyone know the best way to join DOR parcel data with OPA parcel data?”</i>

As stated above, this coding schema was developed following the new H-OGD-I model. After examining the forum posts and interview data, we found no challenge displayed or mentioned in the behaviors of *Evaluation* and *Data Sharing*. Therefore, these two behaviors were excluded from this codebook.

4.7 Study 5: Investigate Fundamental OGD Literacy Capabilities

Study 5 also adopted both forum posts and interview data to investigate fundamental OGD literacy capabilities. Therefore, regarding data collection methods, please see the section of 4.2.2 and 4.2.3.

4.7.1 Data Analysis

After examining all 600 posts, 415 posts were identified that could infer the OGD literacy capabilities. Most of the identified posts are the responses/answers to the initial posts (questions), and some of the initial questions were also selected when the posts included a description that reveal manifest users' OGD literacy. For example, "*I am not able to download 311 requests in CSV format using the provided URL for the last couple of days. It's very inconsistent, is this a known issue? ... Must be a performance issue. I was able to get the data when I filtered out old records using datetime column.*" This post is an initial question, but it presents that this user knows how to download datasets via URL. Ultimately, the 415 posts were further annotated.

The coding schema for OGD literacy capabilities was created based on the new proposed H-OGD-I model, then refined according to the forum posts. The fundamental capabilities were classified in each interacting stage and process. Using the same post example above, we can infer this user is in the process of downloading the dataset. Therefore, this capability is coded into the *Acquire* process.

A student researcher was hired to annotate these posts with me during the content analysis process. This iterative process constitutes five main steps. I first developed an initial

coding schema based on the new proposed H-OGD-I model, and then a first-round random data coding was performed to refine the schema. I also discussed the revised coding schema with my advisors to seek their suggestions. After finalizing the schema, I demonstrated it to the student researcher and trained her on the approach to code the data. Finally, both of us coded all 415 posts. We discussed our coding results every week and eventually reached a 100% agreement.

4.7.2 Coding Schema

Table 3: The Coding Schema for OGD Literacy Capabilities

Behaviors	Actions name	OGD literacy capabilities
Develop data needs	Frame data needs	Be able to determine what data is needed to complete the task. code: Need
Find	Find sources	Be aware of the existence of data and/or knows where to find the needed data. code:Find-Source-ES Be able to identify and locate multiple datasets from the same or different sources to meet the data need. code: Find-Source-MUL
Find	Search with Keywords	Be able to choose keywords or various combinations of keywords to find the needed data. code: Find-KWs
Find	Browse data sub-jects	Be able to leverage the provided categories/groups/tags to look for data. code: Find-Browse
Find	Combine search techniques	Be able to combine search methods (e.g., keyword search + subject browse or keyword search + cited-reference search) to seek data. code:Find-Combine
Evaluate	Evaluate topic relevance	Be able to evaluate if the identified data is relevant to the topic by reading the dataset' title and description. code:Eva-Topic
Scrutinize	Scrutinize usability	Be able to rigorously scrutinize the Usability of the identified data, including identifying file formats, data types and data time period; examining if important data values are missing; checking the accuracy and comprehensiveness of the data; assessing the variables that can combine multiple datasets. code: Scru-Usability
Acquire	Acquire datasets	Be able to download (e.g., provided links) or collect the needed data manually. code: Acq-Manually , or by using programming language, e.g. API, SQL. code: Acq-Program
Continued on next page		

Table 3 – continued from previous page

Behaviors	Action name	OGD literacy capabilities
Understand	Understand context	Be able to know when, where and how the identified datasets were collected, who collected, and what the dataset is about, the scope of the data. code: SenseM-Context-4w1h
Understand	Understand fields	Be able to understand the meaning of column headings and/or the corresponding values by using provided data dictionary or meta-data. code: SenseM-FieldNV-Dictionary ; in terms of personal abilities, such as previous background knowledge, Google search. code: SenseM-FieldNV-Personal
Schematize	Understand table structure	Be able to build connections between the columns, and/or between tables. code: SenseMS-Logic
Schematize	Understand visualizations	Be able to know the basic visualization categories, such as bar chart, heat map, and geographic map, read the information on the Scale, Legend, and Axis, and identify the data format that used in the visualization. code: SenseMS-Visual
Use	Prepare data	Be able to extract, import, and clean the specific needed data through different methods. code: Use-EIC
Use	Prepare (transform) data	Be able to know basic data types, integer, float, string, boolean, character, and know how to convert the data types and transform data between different data formats. code: Use-Trans
Use	Prepare (merge) data	Be able to know the basic concept and rules of merging data, such as identical field, and be able to merge data through different methods. code: Use-Merge
Use	Analyze data	Be able to analyze data by using statistical/mathematic knowledge and methods code: Use-Analyze
Use	Represent results	Be able to present data by creating visualizations. code: Use-Present
Use	Interpret results	Be able to interpret the patterns or analyzing results to by combine context. code: Use-Meaning
Use	Ethically use data	Be aware of the ethical use of data by knowing related laws and policies code: Use-Ethic
Use	Share data	Be able to know how to share data on different platforms, e.g., Google Sheets and Github. code: SHARE

Table 4: OGD Literacy Capabilities in Holistic Process

Holistic process	Actions	OGD literacy capabilities
Overall	Know various context knowledge	Be able to know a broader background information for datasets, including the origin of a dataset, or some stories for the dataset, or some information about data publishers; be able to know other datasets that could be related to the identified datasets or that can meet the data needs. code: Context-Overall
Overall	Locate contact information	Be able to find contact information through Metadata information. code: Contact-Metadata ; be able to find contact information by having related background information. code: Contact-RBI

5.0 Results

This chapter demonstrates findings for answering the RQs. Accordingly, the structure of this chapter follows the RQs. RQ1: What are the typical user behaviors when interacting with OGD? RQ2: What are the challenges users face when they interact with OGD? And RQ3: What are the fundamental OGD literacy capabilities that enable users to use OGD?

5.1 What are the Typical User Behaviors When Interacting with OGD? (RQ1)

To answer RQ1, three sub-studies were conducted. Study 1: observing users' OGD accessing behavioral patterns with a transaction log analysis approach was performed to answer RQ1.2 concerning the user accessing behavior patterns, including the channels users adopted to visit OGD portal websites and behavioral preference of browsing or keyword searching. Study 2: identifying users' individual OGD behaviors with forum post analysis method and Study 3: exploring users' cascading OGD behaviors with interview studies were carried out to understand user task types (R1.1) and typical user behaviors in each interacting stage (RQ1.3).

5.1.1 What Types of Tasks are Performed When Users Interact with OGD? (RQ1.1)

The specific task descriptions and task types can only be identified through the interview method, as the other two studies cannot ask the users further questions. Therefore, the findings for RQ1.1 were obtained from Study 3. In Study 3 (interview analysis), within the 14 interviews (3 group interviews and 11 individual interviews), two primary task types emerged: data-inspired tasks (7 tasks) and problem-inspired tasks (7 tasks).

Data-inspired task refers to a task that starts from data, through exploring existing

data to propose a problem, and then translates the abstract problem to data-centric questions, to finally address the questions by using data. *E.g.*, “We started it by looking at what datasets were available from WPRDC. We just collected some of the ones we thought were interesting or related to each other, and then, kind of, went from there, and we discussed what we might be able to do with them, narrowed them down to the ones that would actually work, and then we use that. So, we did that and then formed the research questions” (Group 1: P1 and P2).

Problem-inspired task refers to a task that starts with the aim of addressing a real problem, then translates the real abstract problem to data-centric questions, and finally addresses the questions by using data. *E.g.*, “Basically, we knew what we wanted, and then, we looked for specific datasets to fill those needs, and then, when we didn’t find exactly what we wanted, we just alternated it a little bit to fit the datasets we did find” (P7).

Table 5 presents the tasks and the task types based on the interviews. Eleven participants (groups 1 & 2 & 3, P3, P6, P11, and P15) started from exploring the existing data, which were identified as data-inspired tasks; they explored the datasets first to discover a topic/problem that is of interest to them and then framed the corresponding data-centric questions. On the other hand, Seven participants with problem-inspired tasks started their tasks by framing data-centric questions in terms of their initial real problems.

We observed that for a data task, even though the task is problem-inspired, the data-centric questions could still be most likely modified to adjust the limitation of the availability of OGD. In addition, these two types of OGD tasks indicate two different starting points for starting a task, which inform OGD portal designers that both supports for browsing and keyword searching datasets are essential to enhance the accessibility of OGD.

Table 5: Tasks and Task Types

Task Types	Tasks	Interviews
Data-Inspired	Answer questions through analyzing data	Group 1 (P1,P2)
Data-Inspired	Answer questions through analyzing data	Group 2 (P4,P5)
Data-Inspired	Answer questions through analyzing data	Group 3 (P8,P9,P10)
Data-Inspired	Answer questions through analyzing data	P3
Data-Inspired	Answer questions through analyzing data	P6
Data-Inspired	Answer questions through analyzing data	P11
Data-Inspired	Teach through using data as examples	P15
Problem-inspired	Answer questions through analyzing data	P7
Problem-inspired	Correct existing database through collecting data	P12
Problem-inspired	Answer questions through analyzing data	P13
Problem-inspired	Create one big shapefile through collecting and using data	P14
Problem-inspired	Answer questions through analyzing data	P16
Problem-inspired	Make personal decision through analyzing data	P17
Problem-inspired	Make personal decision through schematizing data	P18

5.1.2 What are the Users' Accessing Behavioral Patterns When Interacting with OGD and Its Portal Websites? (RQ1.2)

5.1.2.1 Commonly Used Channels for Accessing OGD Portals. Figure 16 shows the channels that users utilized to visit the three OGD portals. Following the terminology in Google Analytics, the classification of the channels shows as follows:

- *Organic Search* refers to the cases when users enter the portal via a search conducted at search engines like Google or Bing.
- *Direct* refers to the cases where a user comes to the portal site directly.
- *Referral* means that the users visit the portal via clicking on a link from another website (i.e., the user enters the portal via a referral).
- *Social* refers to the cases where the users enter from a social media site such as Facebook.
- *Email* refers to the users visiting the portal via clicking on a link from email messages.

- *Others* refers to an acquisition source or medium that is not recognized within Google’s default system defined channel rules like “Email” or “Social”.

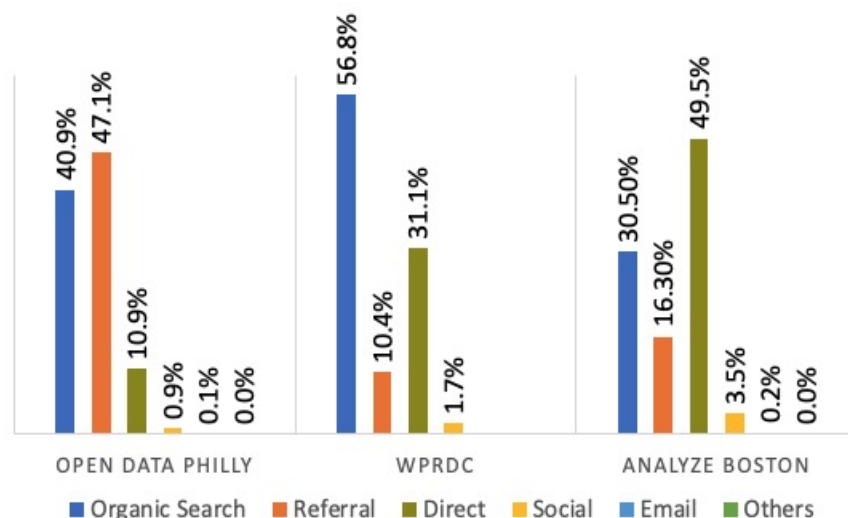


Figure 16: The Distribution of Channels for Accessing the Portals

The results demonstrate that the most used channels in the three different portals are all different from each other. In OpenDataPhilly, the most used channel is referral (36,526; 47.1%), in WPRDC it is organic search (12,403; 56.8%), and in Analyze Boston it is direct (24,095; 49.5%). Particularly notable is that organic search is not the most popular channel for users in OpenDataPhilly and Analyze Boston, which means that rather than directly using search engines, more users visited the two OGD sites via other related sites (OpenDataPhilly) or the URL (Analyze Boston). Only in WPRDC is the organic search the most used channel, which points out that most of WPRDC’s users prefer to locate OGD by using search engines. Also, the only common thing among the three OGD portals is that social media played little impact on engaging potential users. These findings have some differences from general webpage access where both search engines and social media play an essential role in drawing access to certain sites.

These findings might demonstrate that the organizers of OpenDataPhilly have done an excellent job in publicizing the data portal to related communities and the general public. Also, we could assume that more users of Analyze Boston are regular users. They might bookmark the link so that they can simply use the link to access data, or they are really

familiar with the portal and remember the URL. Even though most WPRDC’s users prefer using organic search to others, this finding is still different to Kacprzak et al. (2019)’s results that the majority of users (62.32% for DGU; 74.33% for ONS) find the two portals through web search engines. Their data showed that direct is 14.3% for DGU, 8.52% for ONS, and referral is 9.62% for DGU, 8.52% for ONS.

Commonly used channels for successful accessing OGD portals. The results are not exactly the same when examining successful access cases. Recall that we view a download event as the evidence of successful access. Figure 18,17,19 show the frequency comparison of the channels being used by unique users between the overall access and the successful access when downloading datasets among the three portals. For example, regarding OpenDataPhilly, there are 47.1% (36,526) of overall access cases originated from the channel of referral, whereas the successful access cases only 37.3% (5,097) were from channel referral; with successful access, organic search becomes the most accessible channel.

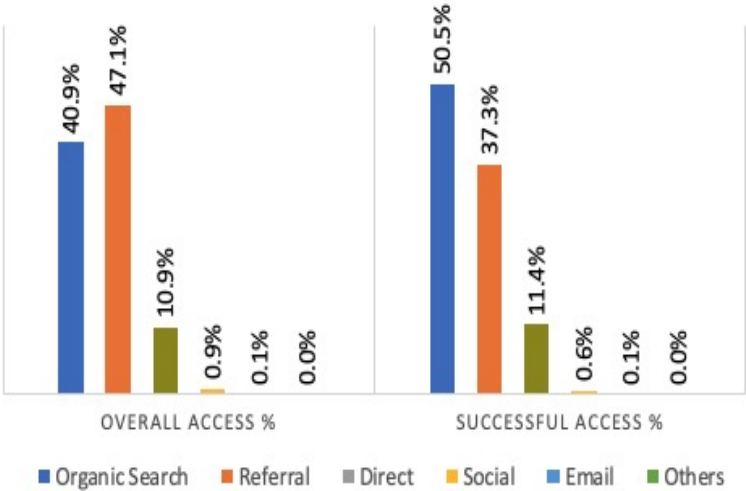


Figure 17: The Distribution of Channels for Accessing the Portals—OpenDataPhilly

Within the context of WPRDC (see Figure 18), its organic search that is the most used channel, and with the successful access, the organic search is still the first one, and the trend is very consistent from overall access and successful access.

Regarding Analyze Boston (see Figure 19), direct wins the most used channels from both overall (24,095; 49.5%) and successful access (1,189; 44.9%), which means that most visits

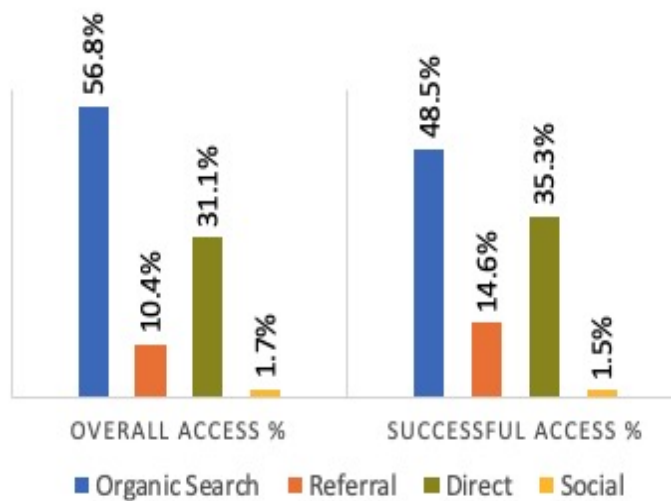


Figure 18: The Distribution of Channels for Accessing the Portals—WPRDC

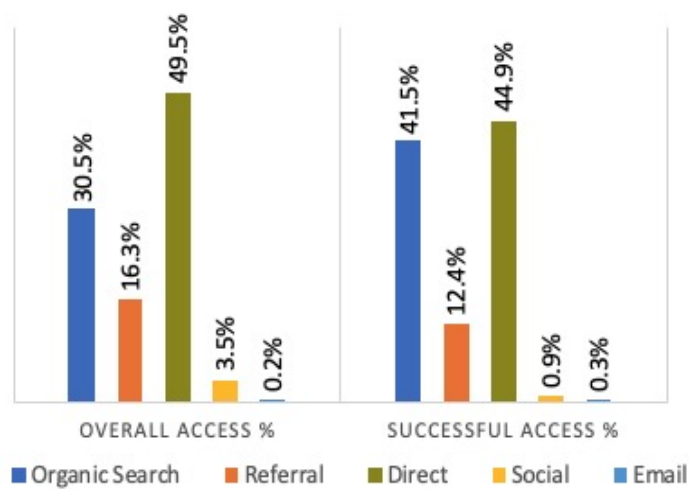


Figure 19: The Distribution of Channels for Accessing the Portals—Analyze Boston

and download events occurred to the scenario of directly typing in Analyze Boston’s URL. Nonetheless, compared to the overall access, the gap between direct and organic search is not that large for the success access, overall (direct: 49.5% vs. organic search: 30.5%) and the successful cases (direct: 44.9% vs. organic search: 41.5%).

It is interesting to see the disparities of the most used channels among the three portals, even though they are all local level portals and the categories of datasets are similar to each other, and they even use the same CKAN platform. We assume that the difference might come from three main aspects. The first is the human side where the various users’ background knowledge, preferences, habits, and tasks could affect their behaviors. The second is the policy aspect, where the different operational ways of the portals. The third one is the interface design of the portals. Although the three portals adopt the same platform, they have diverse interfaces for presenting categories of datasets and the keyword search function. The three aspects form a triangle – users, portals’ operation management, and systems - all of them are important to make OGD more accessed and used. The results also demonstrate that organic search generated more successful access than others in the portals of OpenDataPhilly and WPRDC, and even if the organic search is not the most popular one in Analyze Boston, it still has a high rate for the successful access. This finding may indicate that search engines help to pin down to some specific datasets that the users want, and users’ own knowledge on the portal helps them to locate the needed datasets.

The conversion rate for each channel among the three portals. The conversion rate indicates how much access is converted to successful access from one channel, and the formula is $\text{Conversion rate} = \frac{\text{the number of successful access}}{\text{overall access}}$. For example, in OpenDataPhilly, the overall access from organic search is 31,722, and the successful access through organic search is 6,900, so the conversion rate $= 6,900/31,722 = 21.8\%$. Figure 20 shows the total conversion rates of all channels among the three OGD portals. An obvious observation of this data shows that the conversion rates of all the channels in OpenDataPhilly are much higher than the other two portals. The conversion rates in WPRDC and Analyze Boston are only single digits, which means many more users from OpenDataPhilly downloaded datasets than the other two portals. Only looking at this data would not be able to explain why there is this big difference. However, the subsequent semi-structured inter-

views revealed the hidden reason, which is highly likely caused by the dysfunctional dataset preview function on the OpenDataPhilly website. The users of OpenDataPhilly stated that since the preview function did not work well, they had to download the data first to inspect it. In contrast, on other portals' websites, users may take advantage of the preview function to examine whether or not the identified dataset is the needed data before downloading the dataset. WPRDC's users did not complain about the preview function.



Figure 20: The Total Conversion Rates of Portals

Figure 21 presents the conversion rates of each channel of the three portals. Organic search is the highest conversion rate to Analyze Boston, which is different from previous results – direct is the most used channel and also the channel with the most downloaded datasets. For WPRDC, organic search is the most popular channel and also the channel with the most downloaded datasets. However, the conversion rate is lower than referral and direct. This could be explained as WPRDC may not widely publicize their data to the public, so most users use search engines to obtain their data, yet they have focused on communicating or providing services to certain groups of people or communities that involve many potential users of WPRDC. Considering OpenDataPhilly, the referral is the most used channel, but when we see the successful access, the referral is down to second place, and when we see the conversion rate, referral keeps dropping to the third. It might because referral from the related sites could tell people there are existing datasets but lack clear directions to the

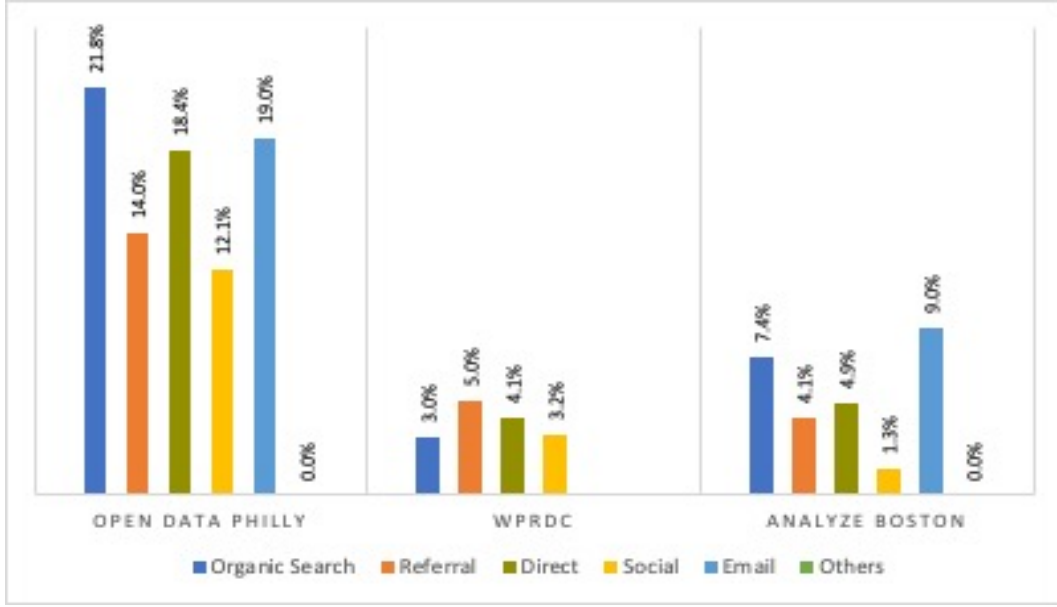


Figure 21: The Conversion Rates of Each Channel Among Portals

needed datasets. Or, people may just look at a website and encounter the data accidentally; they may access the portal due to interests or curiosities. However, in this scenario, the individuals may not download a dataset.

5.1.2.2 User Preferences on Browsing and Searching. In this dissertation, the *Browsing* behaviors are identified as a set of intentional seeking activities on online systems (the goal could be clear or vague), where these seeking activities include the use of the navigations or filters provided by the platforms rather than forming their own keywords/queries to search the needed information. Also, the *Keyword searching* behaviors are considered to be when individuals are able to form queries, enter a query on the OGD portals, and perform either breadth-first move (exploratory) or depth-first move. They may keep refining their queries to achieve the final goals. The main difference between keyword searching and browsing is that keyword searching requires users to know the appropriate keywords for searching, while the browsing behavior needs users to familiarize themselves with the interface design and the infrastructure of the datasets, including the navigations or components of a dataset.

When the keyword searching behaviors trigger the navigation or filters, or vice versa, we consider it as a combination of browsing and searching.

Figure 22 presents users’ accessing behaviors within the three OGD portals. The data demonstrate that more OGD users chose to use browsing (63.4%, 60.0%, and 62.2%) to find the needed data over the keyword searching (33.7%, 35.1%, and 35.0%). All three portals’ users are found to prefer browsing over searching in accessing the datasets, which is contrary to our common impression that more people intend to issue queries to search rather than slowly browsing. We found some explanations according to interview participant descriptions, which are discussed in the 6.1 section.

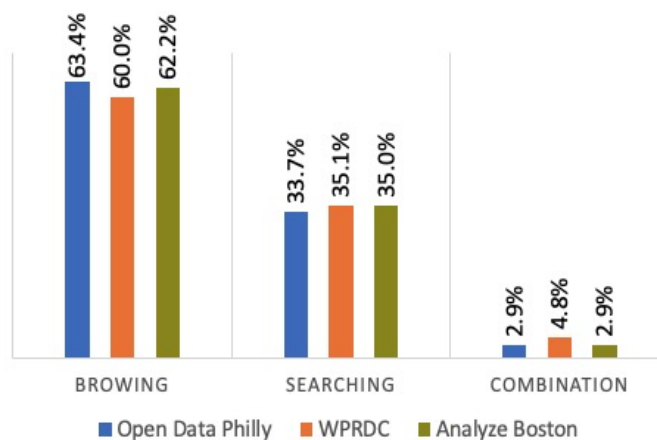


Figure 22: The Accessing Behaviors among the Three OGD Portals

5.1.2.3 User Preferences on Data Entries. All three OGD portals are built on the CKAN platform and have similar classifications of browsing navigation. The following are the explanations for each classification.

Organization. This refers to all the organizations or public sectors that publish their datasets on the platform of the portals. Under this category, users can look for datasets by accessing the organization that could be the publisher of the datasets.

Topics/groups. The topics/groups are classified as different topics, such as health and recreation. Users can find the needed data by going through these topics.

Tags. Data publishers added the tags to the datasets when preparing to submit the datasets to the portals. In addition to classifying the datasets into the topics, adding the

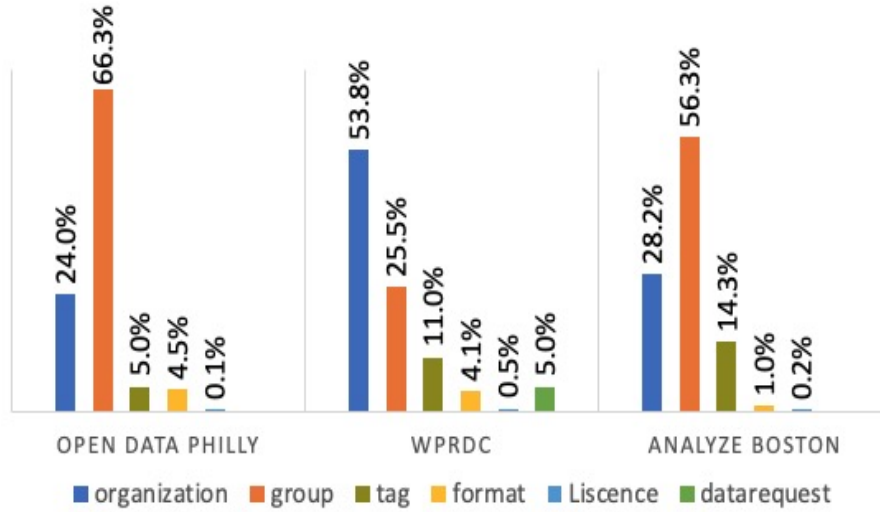


Figure 23: The Data Entries in the Three OGD Portals

tags could provide a more comprehensive description for the datasets. For instance, the dataset “Allegheny County Older Housing” is under the group of Housing Properties; its tags include ACHD DADH, housing, housing age, social determinants. Besides browsing by those tags, when users input those queries to the search box, the dataset would be presented on the result page.

Format. One dataset could have many different formats, including pdf, CSV, Html, etc.

License. Users can browse the datasets based on who provides the license to the datasets.

Data request. Data requests indicate that if users want some specific data and are not able to find it among all the published datasets, users can create a new data request specifying the data that they want to get. Currently, only WPRDC has this entry that allows users to browse data from data requests.

All these classifications can be worked together, like multi-filter functions. Figure 23 shows the users’ preferences for browsing entries. We found organizations and topics are the most used entries for accessing entries, specifically, most users of OpenDataPhilly and Analyze Boston prefer topics to organizations, while most users of WPRDC like to use organizations as the entry. This result makes more sense to us because, generally, when

people are looking for data, they know what topics the data will include as opposed to not necessarily knowing which organizations the data belongs to. Given this finding, we can notice that the accuracy of the topic classifications is very crucial; it can guide users to effectively find what they are looking for. “Organization” is the most commonly used entry for WPRDC and the second one for OpenDataPhilly and Analyze Boston, which makes us infer that many OGD users may be data professionals or working closely with governmental organizations, since this user population would be more clear about the infrastructures of the organizations and datasets published by these corresponding organizations.

5.1.3 What are the Typical User Behaviors in Each Interacting Stage? (RQ1.3)

According to the forum posts (FP) analysis (Study 2) and the interview data (Study 3), five data interaction stages were identified: task preparation, data foraging, data sensemaking, data use, and data sharing. The corresponding user behaviors were pinpointed and also classified (see Table 6).

Table 6: Typical User Behaviors Identified by the Analysis of Forum Posts and Interview Data

Stages	User behaviors	Frequency (FP)	%	Frequency (interview)	%
Task Preparation	Explore existing data	N/A	N/A	11/18	61.1%
Task Preparation	Frame data-centric questions	N/A	N/A	18/18	100%
Task Preparation	Develop data need	N/A	N/A	18/18	100%
Data Forage	Find	94/292	32.2%	18/18	100%
Data Forage	Evaluate	N/A	N/A	18/18	100%
Data Forage	Scrutinize	54/292	18.5%	18/18	100%
Data Forage	Acquire	64/292	21.9%	16/18	88.9%
Data Sensemaking	understand	32/292	11.0%	18/18	100%
Data Sensemaking	Schematize	11/292	3.8%	16/18	88.9%
Data Use	Prepare	16/292	5.5%	16/18	88.9%
Data Use	Analyze	6/292	2.1%	14/18	77.8%
Data Use	Represent	1/292	0.3%	17/18	94.4%
Data Use	Interpret	1/292	0.3%	14/18	77.8%
Data Use	Communicate	3/292	1.0%	15/18	83.3%
Data Use	Make decision	1/292	0.3%	2/18	11.1%
Data Share	Share	13/292	3.1%	1/18	5.6%

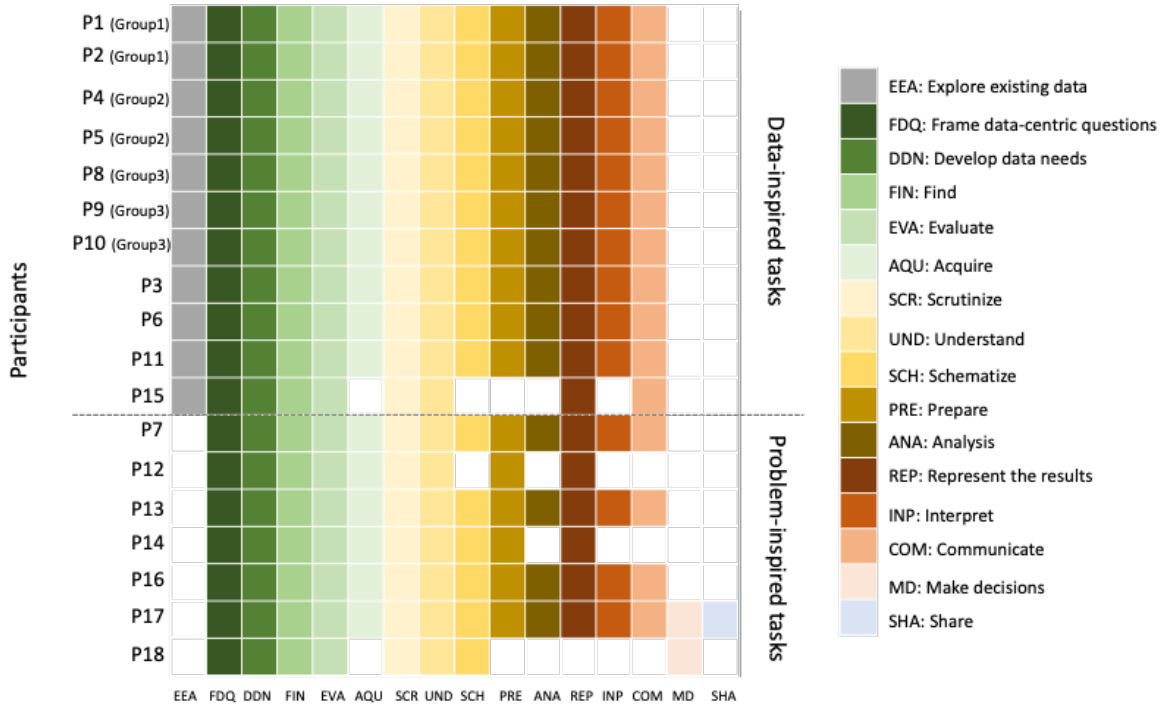


Figure 24: Behaviors Manifested by Each Participant

An overall observation from Study 2 (forum posts analysis) indicates three OGD interacting behaviors could not be identified through forum posts data analysis. Referencing Table 6, the behaviors of *Exploring existing data* and *Framing data-centric questions* were not pinpointed by examining the forum posts because as OGD users sought help in the forum, they already clarified their data-centric questions, the help they sought on the forum were further steps than framing data-centric questions, which can not be manifested in the posts. Likewise, the behavior of *Developing data needs* can be inferred from the posts; however, in order not to repetitively count the data, the data from these posts are exclusively recorded under the Finding behavior. Furthermore, we found that not all the behaviors identified in this study are necessary for each task. Figure 2 depicts the sequence of behaviors reported by each interview participant. For example, in a case when the goal of the task is primarily to generate a geographic map (P14), after foraging and schematizing all the collected data, the participant only needs to create and represent the map; the behaviors of analyzing, interpreting and making decisions may not be required while performing the task.

5.1.3.1 Task Preparation Stage. The task preparation stage denotes the process of coming up with a real problem that a task aims to address, translating the real problem into related data-centric questions, then discovering the data needs for answering the questions. The real problem can be proposed by an existing topic, such as work responsibility or personal interests, or by exploring existing datasets to develop a real problem. By the end of this stage, users have developed data-centric questions and data needs. Below are the three primary behaviors (Figure 25) associated with this stage.

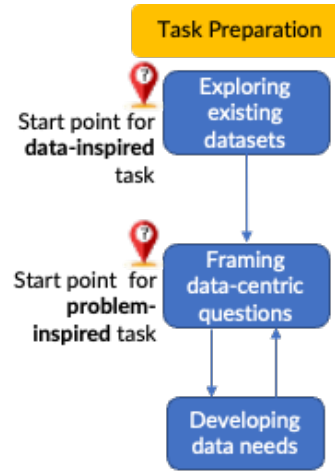


Figure 25: H-OGD-I Model — Task Preparation Stage

Exploring existing data denotes the process when users try to propose a real problem that interests them by using the existing datasets. The typical behaviors include browsing the categories or tags of datasets and evaluating the topic’s relevance. *Exploring existing datasets* here is the starting point for a data-inspired task. E.g., “I just went to WPRDC; they have the left-hand columns that describe what types of data they have. I was interested in environmental and city management data, so I clicked both and browsed through the list” (P1).

Framing data-centric question(s) refers to translating an abstract problem into data-centric questions to address the goal of a task. *Framing data-centric questions* is the starting point for a problem-inspired task. E.g., “How many unique instances of polling place location change occurred in each municipality over the three years reviewed?” (P3).

Developing data needs denotes probing into what data are needed to answer the questions proposed by the task, then developing a clear list of required data. E.g., “*Is the history of COVID cases by zip code available? We are using the data to analyze the impact of the pandemic in our city, and the historical case count by date and zip code would be very helpful to understand local trends*” (FP).

As observed, framing *data-centric questions* and developing *data needs* evolve as users interact with data or learn new information. For instance, when users could not find the exact data they needed, they went back to tweak the data-centric questions to make the questions answerable. Or, the analysis results may inspire users to propose more intriguing questions. Therefore, *Framing data-centric questions* and *Developing data needs* are dynamic processes.

5.1.3.2 Data Forage Stage. When data needs are generated, all interviewees (18) described that they experienced the process of data foraging, including the behaviors of finding, evaluating, scrutinizing, and/or acquiring. Study 2 (FP analysis) also supports this finding, presented in Table 6. By the end of this stage, users have found relevant and usable datasets. Below are the four primary behaviors (Figure 26) associated with this stage.

Finding denotes the process when users look for data, including identifying data sources and applying data-seeking strategies to find data. The three detailed facets of *Finding* behaviors – channel, keyword searching, and browsing that were proposed by the initial H-OGD-I model were also revealed by Study 1 (transaction log analysis) and Study 3 (interview data analysis). According to Study 1 (transaction log analysis), a channel indicates a source or manner that assists people in finding the OGD portal’s website. We found that OGD users use various channels to access OGD portals, such as search engines and referrals. Study 3 (interview data analysis) made the same discovery (e.g., P12, P14), which also aligns with previous work (L. M. Koesten, Kacprzak, et al., 2019). Moreover, Study 1 (transaction log analysis) revealed that OGD users use the strategy of browsing, keyword searching, or combining these two methods to seek data. For further exploration, we observed that OGD users prefer browsing data over using keywords to find data. In Study 3 (interview data analysis), the participants described that they used keyword searching and/or browsing

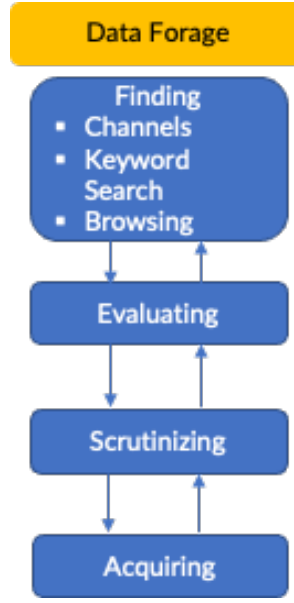


Figure 26: H-OGD-I Model — Data Forage Stage

strategies to find information. For example, P5 searched with keywords: *“we did it for the search bar; we just started typing in the ‘voting data,’ and then one popped up that was ‘polling places.’* P7 used filters on the OGD platform to browse: *“I remember there were different tags on the left hand, so I used those to try to filter it down a little bit”*.

Once participants located the data, they began to evaluate the returning results. ***Evaluating*** denotes the first look when users examine the topic relevance of the identified dataset with their data needs. The actions could involve reading the title and description of the dataset. All the participants (18) stated that they evaluated the relevance of a dataset to their data needs. E.g., *“The first look is reading the topic and description of the dataset. On WPRDC, they give the descriptions of, what you are going to look at, what year, and the available file formats”* (P4). Forum posts data cannot identify the *Evaluating* behaviors, and we assume that it is because users would not ask questions about whether or not the data is topic-relevant to their needs.

After confirming the identified data is topic-relevant, interviewees further explored if the data is usable for their task. ***Scrutinizing*** denotes the second look when users rigorously scrutinize the utility of the identified dataset, including examining the file formats, specific

data types, data time, missing data, and the accuracy of the data, assessing the variables that can be used to combine multiple datasets. In Study 3, all interviewees (18) described the behaviors of scrutinizing data. E.g., *“there are errors within the testing positive and negative cases. They sometimes retracted from one day to another...”*(P13). Moreover, Study 2 (FP analysis) reinforces this finding; *Scrutinizing* is the third most popular question in the posts (54, 18.5%). Specifically, users focused on examining four aspects of data, as shown in Table 7. To clarify, as one post could inspect multiple aspects, the sum of the frequency (56) in Table 7 is higher than the sum (54) in Table 6.

Table 7: Four Aspects that OGD Users Mainly Scrutinized

Scrutinize aspects	Explanations	Frequency
Missing data	Examine the data fields and values. The dataset topic is relevant, but the needed data fields are missing. Or, the data fields exist, but the corresponding values are missing without any explanation. E.g. <i>“When I opened the 2016 csv file, I couldn’t see any data other than object ID. I am particularly interested in tree species.”</i>	21
Time	Examine the specific time range covered by the datasets. E.g., <i>“I used to use this dataset for real estate analysis and research. I see the most recent sale dates end in 2020. From the activity log, looks like the last update was 10-11 months ago. Does anyone know how this dataset can be updated?”</i>	21
Content	Examine the specific content of a dataset, namely the scope of a dataset. E.g., <i>“Does anyone know if there is a possibility that older crime incidents are now being added to the dataset?”</i>	8
Accuracy	Examine the accuracy of a dataset, namely sparse and duplicated data. E.g., <i>“I have been eyeing the tracking of COVID death data per date and noticed that there is increasingly sparser data the last few weeks and repeat values.”</i>	6

Acquiring denotes the process when users obtain the datasets. Study 1 (transaction log analysis) identified the downloading behavior in terms of the downloading times. In addition, sixteen out of 18 interviewees claimed they downloaded datasets at least once (Study 3). E.g., *“we downloaded it first because it was just fragmented when you just looked at it (P3).”* Based on the forum posts data analysis (Study 2), the *Acquiring* behaviors in-

volve clicking the provided link to download the datasets directly (44 out of 64), *e.g.*, “Hello, I am trying to download CityBasemap. But when I click on to download it takes me to the services directory page” (FP), or querying data through APIs (13 out of 64), *e.g.*, “I have not seen any new data beyond 4/1/2020 from the permit download API URL at: https://phl.carto.com/api/v2/sql?q=SELECT * FROM permits. Is this still the valid permit API URL or is there a new one that should be used?” (FP).

Other less-used ways of acquiring data are manually copying and pasting data (3 out of 64), using web applications (1 out of 64), and making an online request (1 out of 64). Several cases also show that users employ a combination of the acquiring behaviors. For instance, after noticing the downloading link did not work, the user tried to use APIs (1 out of 64) or web applications (1 out of 64) to access data.

Our initial conceptual model did not include the behavior of *Acquiring* data. Many conventional information-seeking behavior models do not account for the fact that downloading datasets is a critical characteristic of using data rather than using information. This was demonstrated when analyzing forum posts, where “how to download a dataset” was the top two asked question (64, 23%). Generally, if someone wishes to use data, they must download it in order to manipulate and use the data, while with information, users can directly read it and understand it in order to use it. This *acquiring* behavior can occur either before or after closely examining the data, depending on the user’s preference for the data preview function and whether the preview function is functioning properly.

5.1.3.3 Data Sensemaking Stage. Once data were obtained, participants began to make sense of the identified data, including the behaviors of *Understanding* and *Schematizing* (Figure 27). By the end of this stage, users had gained a better understanding of the data, and had constructed meaning from it. For Study 3, 18 participants reported that they had an *understanding* behaviors, while 16 of them stated that they schematized the data to gain a more in-depth understanding of the data. Study 2 revealed that 42 posts indicated the users were in the *Data Sensemaking* stage. Of these posts, 32 were attributed to the *Understanding* behaviors, and 11 were aligned with the *Schematizing* behaviors. Additionally, one post covered both *Understanding* and *Schematizing* behaviors.

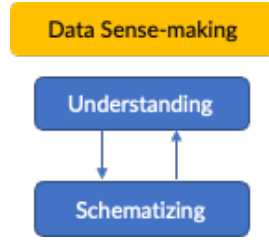


Figure 27: H-OGD-I Model — Data Sensemaking Stage

Understanding denotes the process when users start to perceive the intended meaning of the data, such as understanding the meaning of the column headings and the values of the data. Typical behaviors include closely reading the data dictionaries and/or “read me” files. E.g., *“I want to say that for most of them anyway, the field names kind of match what I would expect them to be, but there was a couple that, like it could be one of two kinds of things, especially, for things like how large areas are? Unless it really specifically says 8 acres. I had to consult at least two data dictionaries that were included. I did use a data dictionary to make sure we used the right data” (P2). “Does anyone know how they collect the information on streetsmartphl?” (FP).*

Most OGD is structured data which requires contextual information to explain the fields and values. Hence, a great number of OGD portals currently provide additional documents, e.g., data dictionaries, for describing the data, and users benefit from reading the documents to understand the meaning of the data.

Schematizing denotes the process when users try to establish the logical relationships between the data fields or data records, namely, organizing the data with relevant information or building connections between the data. Typical behaviors include cottoning on the content of data and/or the visualizations provided by platforms and manipulating data, such as sorting, to find connections between data within one dataset or multiple datasets. E.g., *“The mortality data looks like through 6/10/20 we have a clinical date of death as far as 6/2/20 (with a total of 1,385 incl. NA for 10), while the death by zip and by sex/age tallies have totals of 1,421 and 1,428 deaths, respectively. . .”(FP).*

As mentioned, 16 out of 18 interview participants described their schematizing behavior. Regarding the two tasks without the schematizing behaviors, the goal of one task is to correct the inaccurate data stored in their organization. Therefore, the main work was collecting related data from the OGD portal, then revising and storing them in their database. This participant was still required to understand data to ensure she located the correct data, but there was no need to further schematize the data. Additionally, the purpose of the other task is to guide students in creating data explanation documents for OGD, so the main work of the instructor is to facilitate students to locate and understand data. Therefore, this task may exclude the schematizing process.

5.1.3.4 Data Use Stage. Ultimately, users begin to use data to implement their tasks. In terms of our data, all participants (18) depicted their ways of using the collected data (Study 3); the forum posts data also demonstrated the behaviors of using data (Study 2). The specific examples are given in following corresponding sections.

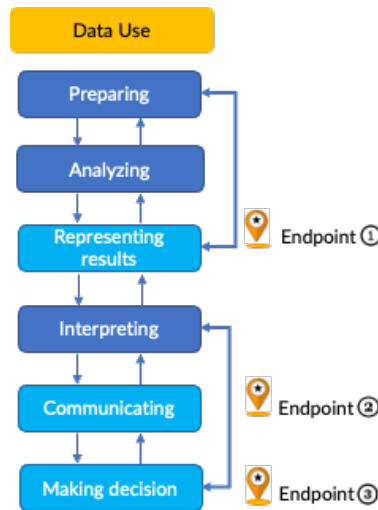


Figure 28: H-OGD-I Model — Data Use Stage

The stage of *Data use* denotes the process when people apply the collected data to perform their tasks. The possible activities include preparing data (e.g., cleaning, transforming, and merging), analyzing data, representing results, interpreting, communicating with data, and making data-driven decisions.

Preparing refers to the process of making the data ready to be further analyzed, visualized, or directly used (e.g., building a database). The specific activities could involve cleaning, extracting, importing, transforming, and merging. E.g., *Cleaning*: “I think that was the green space dataset, which had a lot of tiny green spaces that we wanted to take out” (P1). *Importing*: “Anyone have a quick instruction on importing or linking to maps showing police districts for Philadelphia? I have tried different approaches, but I haven’t gotten far.” (FP).

Analyzing denotes the process of utilizing various methods (mainly quantitative) to analyze data to discover useful information, identify patterns, draw conclusions, and/or support decision-making. E.g., “Now I have the table data for 2019 along with the field descriptions. I am not understanding how to correlate the location data i.e., how to get locations for that data. I have the census_block_group as a column in the data set but that’s about it. I also read about Geo_IDS. I am now confused how to proceed.” (FP).

Representing results depicts the process of representing the results through different methods. According to our data, the primary methods are visualization and statistical analysis results. E.g., “Being able to understand what the number meant for our project. Looking at this number, this is the percentage rate of change that happened between polling places from 2017 to 2020 in this municipality in Allegheny County. We also used visualization, so we created a graph, and just being able to look at that graph and being able to explain it” (P5).

Interpreting indicates the process of exploring and translating the analysis results through combining the comprehension of domain knowledge and the data analysis results to finally draw conclusions or make decisions. This process translates the data-centric questions back to knowledge or information. E.g., “I think that people see things like graphs and charts that kind of stuff all the time, everywhere, so knowing what is behind that, knowing that you are able to go and see the source, this is what they showing me out of what is available kind of thing, I mean a lot of time that stuff is showing what an author is intending to show which may or may not be the entire picture.” (P2).

Communicating manifests the process of using appropriate language and manners to communicate the data with targeted audiences, such as telling an evidence-based data story.

E.g., “our difficulty was using the vocabulary to describe what we were seeing and talking about; we wanted not to say anything bad.” (P5).

Making decisions describes the process of making a choice or deciding to take action based on understanding the identified data or the interpretations of the analysis results. E.g., “I used the government data to see the safety of the neighborhoods in Pittsburgh; we considered moving to neighborhoods with fewer criminal incidents.” (P18).

Significantly, in addition to the behaviors in the stage of *data use*, Study 3 (interview analysis) also revealed three endpoints that describe the scenarios in which users complete their tasks.

Endpoint 1 (Representing results) is the ending point of the task that aims to present the results from OGD interaction. For example, collecting needed data for creating a database, after going through tasking preparation, data forage, and data sensemaking stages, the user prepares the data based on pre-defined requirements and then creates a database to represent the results. The other processes in the *Data use* stage are unneeded (P12, P14, P15).

Endpoint 2 (Communicating) is the ending point of the task that aims to generate a report for someone else; they do not need to make decisions by themselves, for example, students’ assignments or work responsibilities (Group1&2&3, P3, P6, P11, P7, P13, P16).

Endpoint 3 (Making decision) is the ending point of the task that aims to make a final decision by users themselves after interacting with OGD. The process can go through all the stages (P17) or in the middle of some behaviors (P18). For example, a user decides to buy a house in certain neighborhoods after schematizing a crime data visualization in an OGD platform (*Schematizing to Making decision*).

The initial conceptual model did not discuss these granular behaviors and the three endpoints in the *Data use* stage. We discovered them with the grounded theory method by investigating the forum posts and interview data.

5.1.3.5 Data Share Stage. The *Data share* stage was also not discussed in the initial concept model; we disclosed it during the forum post exploration. This stage denotes the process when users share the data collected from various data sources through different

platforms and methods, e.g., GitHub, Google sheets, and/or emails. This behavior stresses that data sharing happens between users rather than data portals or organizations. The results achieved in this stage heavily depend on each user’s needs or preferences, so it is an additional stage of the Model. E.g., *“I started additional cron jobs to cover all the datasets. I’ll start figuring out how to publish them to Github in an organized manner.” (FP).*

5.1.4 The Evaluation of the Behavior Model for Human OGD Interaction

After the typical user behaviors of interacting with OGD were discovered, we built a new model for H-OGD-I. To better validate and refine this new model, the Delphi Method, which aims to “obtain the most reliable consensus of opinion of a group of experts” (Dalkey & Helmer, 1963), was employed. Also, in order to deeply understand experts’ opinions, we used the interview method to discuss with three experts. The first expert was a faculty member in a university in Europe with expertise in both HDI and OGD fields. The second expert has been working for a local OGD portal for more than six years, having rich experience in working with OGD users. The third one is a faculty member in a University in the United States who is an expert in both human information interaction (HII) and OGD fields and has been working on HII for over 20 years.

Three primary criteria were applied to guide this evaluation: understandability, comprehensiveness, and reasonableness. Specifically, understandability refers to whether or not the model’s illustration and visualization are easy to understand. Comprehensiveness denotes whether or not the components of the model are comprehensive. Reasonableness indicates whether or not the concepts for each behavior make sense to the experts. The model was refined after each interview, which means the model that the second professional reviewed is an updated version based on the suggestions from the first expert. Both similar and dissimilar suggestions were obtained from the three experts. Until the third one, we found fewer suggestions were brought up, and the main concern was repetitive. Therefore, we considered this new model was assessed sufficiently.

As a result, all three experts agreed that the components of this model are comprehensive. The first expert we interviewed asked us if all of the behavioral components in the model

necessarily happen to all the tasks. And if not, she suggested making a clear statement that emphasizes the combinations of the components could be different based on the tasks, and this model represents the high-level and comprehensive components. According to this idea, we added the description of the behavioral sequence reported by each interview participant (shown in figure 24), three endpoints, and the corresponding explanations.

The experts also proposed suggestions for the concepts or the terms of the behaviors in the model. For instance, we originally named data-inspired tasks data-driven tasks. After explaining this term to the experts, one expert suggested that data-inspired data is more accurate than data-driven.

In addition, even though we claimed this proposed model is a non-linear process, all three experts advised us to make the visualization more obvious to show the iterative relationship between the identified behaviors in the model. Therefore, we adopted double arrows to show the loops between the contiguous behaviors and added the track back arrows between the stages to visualize the loops between the stages. The final model is presented in Figure 42.

5.1.5 Outcome1: The Model for Human OGD Interaction

We synthesized our findings from the proposed initial conceptual model (theoretical base), the three empirical studies, and experts' evaluations and suggestions, to develop and validate this new model for human OGD interaction (see Figure 7). Two significant additions have been incorporated into this new model. First, the stage – “*Task*” is enriched to “*Task Preparation*”, whose components include the task types and disclose the related behaviors according to the different task types. Secondly, based on the findings from analyzing forum posts, the stage *Data sharing* is added to this new H-OGD-I model. The detailed behaviors of this component are also supplemented. More details about this H-OGD-I model can be found in Section 5.1.

As stated, this H-OGD-I model describes the relationships between the behavioral components at a schematic level, where the exact relationship between the behavioral elements is iterative and associated with a particular individual and a particular task. For instance, users may go back to the *Find* process when they *analyze* the collected data because they need

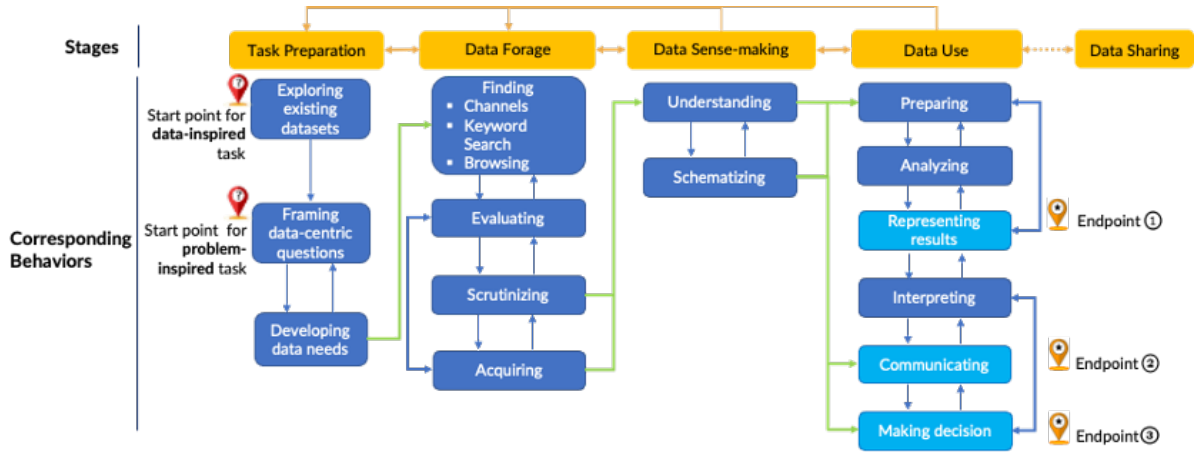


Figure 29: Finalized Model for H-OGD-I

to find another dataset to complete their task. In addition, this H-OGD-I model contains possible behaviors when users interact with OGD, but not all the behavioral components in this model are necessary for every task in practice. The interactions between humans and OGD are nonlinear and dynamic processes.

This model can also be applied to collaborative tasks at an overall level. Three group interviews were conducted in this study, and all the group members described their experiences interacting with data and how they collaborated. We observed that the overall processes are consistent with this H-OGD-I model. Additional collaborating activities include but are not limited to assigning sub-tasks, communicating findings, asking questions, discussing data analysis methods and ways of reporting results, and sharing data with each other. This observation also indicates that the goal of the tasks more decisively decides the behavioral components.

5.2 What are the Contextualized Challenges Users Face in Each H-OGD-I stage? (RQ2)

This section demonstrates the findings of user challenges when they interact with OGD. Based on the new H-OGD-I model, we conducted forum data analysis and interview studies to discover the contextualized difficulties existing in each interactive stage. According to the interpretation of the findings, we also pinpointed the probable cause of the challenges: on the user side, lack of OGD literacy, and on the platform side, insufficient supportive information, or inefficient functionality design. When determining if the challenge is user-side or platform-side, we followed these two principles: when the post descriptions and participant statements mainly focus on the difficulties produced due to lacking data skills, we identified that advancing OGD literacy is a better way to address the difficulties. For example, P1 stated that *“I would say there was one point where I had to add those last two zeros to a bunch of data points... at the very end, I spent so much time trying to figure that out.”* On the other hand, when the post descriptions and participant statements mainly discussed the challenges caused by the platform, we considered that improving the platform functionality design or providing sufficient information is more helpful to overcome the obstacles. E.g., *“I was trying to DL the CSV file, and it kept giving me an error. Can anyone assist?”*

The forum posts were first analyzed based on the new H-OGD-I model that includes the stages of task preparation, data foraging, data sensemaking, data use, and data sharing. In order to identify user challenges, we focused on analyzing posts associated with questions that users posted on this online forum because we considered when users asked questions, they most likely came across challenges. We detected that the challenges mainly happened in the stages of data foraging, data sensemaking, and data use. Figure 31 demonstrates the distribution of the contextualized OGD user challenges according to forum data analysis. In addition, interview data analysis was subsequently conducted to reveal OGD user challenges, and the findings validated and complemented the results from forum posts analysis.

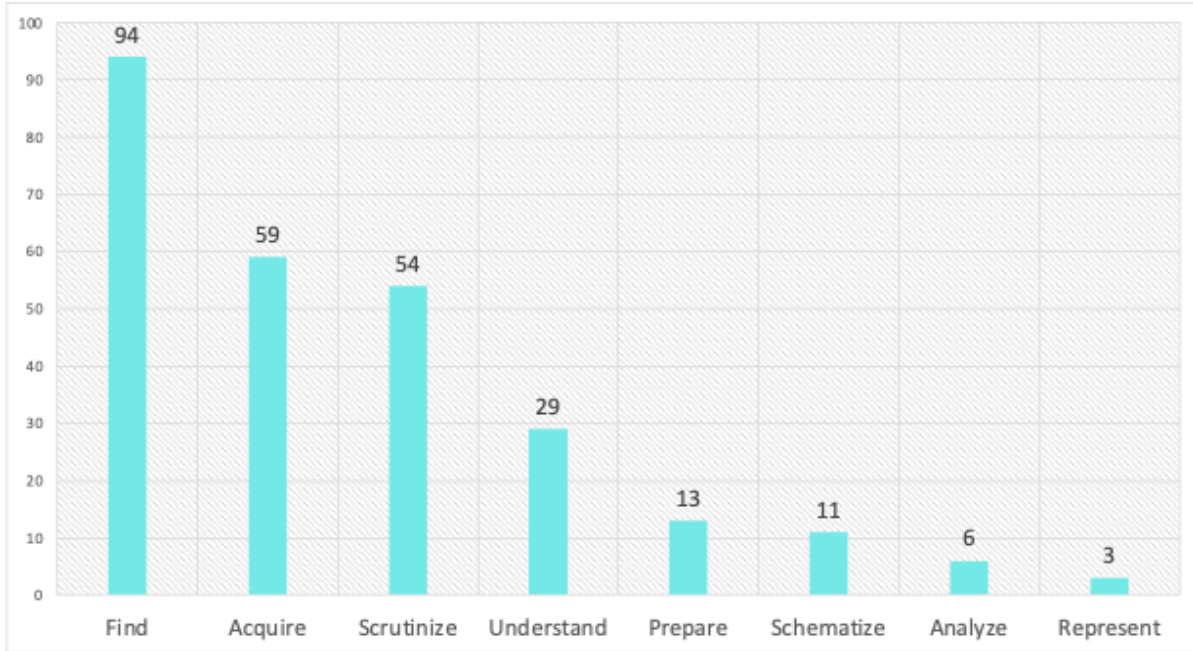


Figure 30: The Contextualized Interactive Challenges User Encountered

Even though we identified the *data sharing* stage in the forum post, no related challenge was found. The forum posts contain the topic of sharing data, but the purpose of these related posts is to inform others under this thread or forum that the user will share the data that they collected from various data sources by themselves rather than asking about the method of sharing data. For instance, “*I’ve set up a google sheet and some scripts to compile the daily open data feed output for each dataset. Not as elegant as posting on GitHub, but it’s available as well to anyone who wants it.*” Also, as discussed in the previous section, the stage of task preparation cannot be observed from forum posts, so the corresponding challenges in the task preparation stage cannot be identified either. However, since we can ask further questions to the interview participants, the challenges in task preparation were revealed by interview data analysis.

5.2.1 The Challenges in the Context of the *Task Preparation Stage*

5.2.1.1 The Challenges in the Context of the *Framing Data-centric Question Process*. Two groups of users (groups 2 and 3) proposed the challenges of *framing data-*

centric questions in the task preparation stage. Group 2 (P4 and P5) described that “*I think the majority of our challenges were when we were trying to figure out our questions, we got all these data, we all thought it was going so well, and then we had the meeting with professors, and they were like it doesn’t really make sense, so then we had to scramble...*” This challenge was more reliant on user OGD literacy skills. In addition, Group 3 (P8, P9, P10) also experienced a hard time in the process of adjusting the data-centric questions and data needs. However, their case was different from Group 2. They stated that framing the questions was a pain. Specifically, after developing the questions and corresponding data needs, they found they could not find all the needed data. Thereby, they had to go back to adjust the questions and seek data again. Even though this group considered this challenge was about framing data-centric questions, the real cause of the difficulty was in *finding* data.

Overall, the interview participants thought there were fewer challenges in the *task preparation* stage, especially for the tasks with a simple goal. For example, the purpose of the task is to examine the safety of the participant’s neighborhood (P18). The data-centric questions could be how many and what types of criminal incidents happened in the community and the comparison with other neighborhoods. Therefore, the corresponding data needs would only be the data on criminal incidents in certain areas, which was easy in the course of task preparation.

In the stage of *task preparation*, the processes of *framing data-centric questions* and *developing data needs* require a certain level of OGD literacy, such as computational thinking and civic data domain knowledge. Therefore, improving users’ OGD literacy will be more effective in addressing these difficulties.

5.2.2 The Challenges in the Context of the *Data Foraging Stage*

5.2.2.1 The Challenges in the Context of the *Finding Process*.

The forum data analysis shows that the most difficult challenge of interacting with OGD is *finding* data (94, 35%). After further examining the questions concerning *finding* data, five main sub-categories of the challenges were distinct, shown in figure 31.

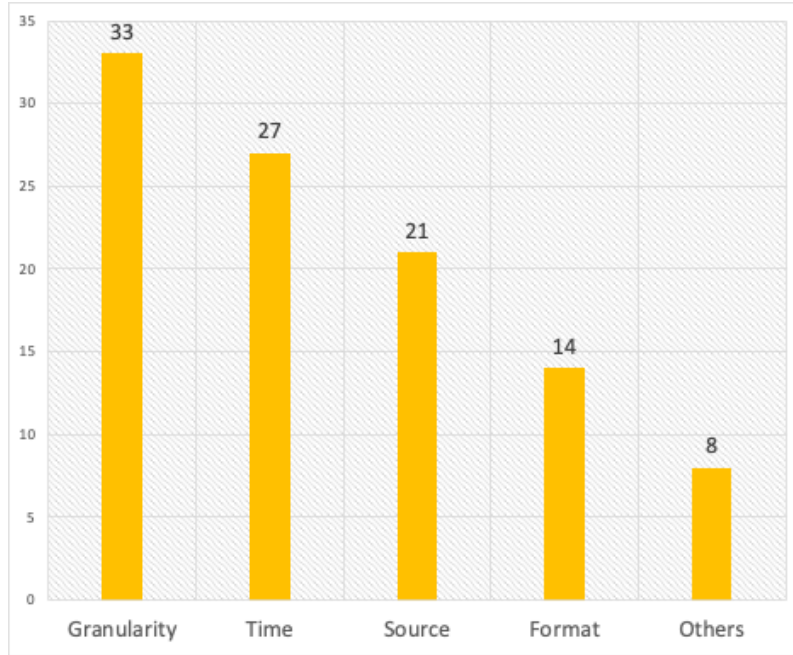


Figure 31: The Sub-categories of *Find* Challenges

The top finding challenge is *data granularity* (33, 32%), which refers to the users knowing the topic of the needed dataset existed and even knowing the possible sources, but only seeking the data with a specific data point requirement, such as explicit data fields and different data dimensions. The post explicitly points out the granularity requirements. E.g., “*I looked through the sites you mentioned, but I was not able to find what I am looking for. I am looking for: - the number of registered voters per ward in 2018.*”

The second most asked question is *data time* (27, 26%), which indicates the users knew the topic of the needed dataset existed and even knew the possible sources, but only sought the data with a specific temporal requirement. The post explicitly points out the temporal requirements. E.g., “*I am looking for property assessment data from 2019. Does anyone know where I can access past datasets for assessment data?* ”

The third sub-challenge is *data source* (21, 20%), which indicates that the users knew the needed topic of data and asked for the corresponding data sources. Regarding *data source*, we observed two levels of source finding. The first level refers to a user not knowing whether a dataset exists. E.g., “*I am looking for a list of all of Philly city’s public parks. Does this*

exist?” The second level refers to knowing the existence of the dataset but having no clue about the sources. E.g., *“Does anyone know if there is a probate list on this site or where I can get it from?”*

The fourth challenging barrier is *data format* (14, 14%), which signals the users knew the topic of the needed dataset existed and even knew the possible sources, but only sought the data with a specific format need. The post explicitly points out the format requirements. E.g., *“Would love to have a file of the city council districts in KML format to upload to google maps so I can create a custom map. Any chance this already exists somewhere? I couldn’t find on the open data website.”*

In addition to the clear classification of users finding challenges, users also looked for other types of information (8, 8%). E.g., *“Office ”29XXXXXX” has had 32 complaints in the last 5 years. Is there a way to find out who officer ”29XXXXXX” is? Would that require a FOIA request?”*

The interview data analysis complements the findings from the forum posts analysis. The questions on the forum posts were straightforward and referenced a specific point and time, while interview participants described their experience of seeking data with a continuous process and overall emotion. For example, P7 states that *“I think the biggest challenge was the range of the search results, so even if I was pretty specific with my keywords, I was getting a pretty wide range of results, things that weren’t in Pittsburgh, even if ‘Pittsburgh’ was like in the title.”* A similar description was that *“they take you to a lot of information, and a lot of charts, but most of it might not be what you are looking for; a lot of it is pretty general.”* (P17).

Both participants (P7 and P17) considered a wide range of search results very challenging, which indicates a better recommendation system could be more helpful to users. In addition, the discovered sub-challenges of *finding* OGD also advise us that even though OGD users found the data sources, they may still be unable to identify the needed data. That is because the most challenging difficulties are that OGD users need granular data or specific time and formats, which are hard to search on the portal’s website, or the dataset description is unclear. In the meantime, the related OGD literacy skills, namely the knowledge of civic data and framing search queries are also essential to address these challenges.

5.2.2.2 The Challenges in the Context of the *Acquiring Process*. The second big challenge for users is *acquiring* data (59, 22%), which surprised us on a certain level. We would not anticipate that downloading data could be complicated.

Based on the forum post analysis, 45 out of 59 questions (75%) were regarding the downloading links. E.g., *“I was trying to DL the CSV file, and it kept giving me an error. Can anyone assist?”* We also found that most posts are complaining about dysfunctional links. According to the responses, we confirmed that most of the challenges are caused by the poor functions provided by the platform rather than users’ ability to download data.

Another sub-challenge of *acquiring* data is using APIs to obtain a dataset (14 out of 59, 22%). E.g., *“Could you explain how to use the API to access the databases? There’s another one in OpenDataPhilly that gets stuck downloading the CSV.”* Since using APIs requires a certain level of users’ OGD literacy, and after examining related responses, we determined that most of the challenges are caused by users lacking the ability to use APIs to obtain data rather than platform services. In fact, most OGD platforms, including the three research settings, provide an APIs method to facilitate large-size data downloading. In addition, compared to 75% of posts asking about downloading links, only 22% of questions regarded APIs. This observation does not confirm that there are fewer users who have challenges using APIs because most users have no idea how to use APIs to download data. Therefore, most users would not ask pertinent questions. E.g., *“I’m not sure about using API, so I’d be grateful to receive one of those CSV links when they become available.”*

The interview data analysis displays similar results concerning the challenge of using APIs to acquire data. For example, for obtaining granular data, P1 described that *“it was hard to find the geo coordinates data because there were lots of different APIs that were pulling from different sources. So, I didn’t choose based on the data itself; I chose based on how easy the API was to use and how efficient it was.”* Regarding downloading links, the participants who used WPRDC’s data considered the function worked very well, while the participants who were OpenDataPhilly users thought this function needed improvement. This observation indicates that the functionality offered by OGD platforms plays an essential role in better serving users.

Overall, for acquiring data, sufficient functions provided by OGD portals (offering func-

tional downloading links) and OGD literacy are both needed (using APIs).

5.2.2.3 The Challenges in the Context of the *Scrutinizing Process*. The third top challenge of interacting with OGD occurred in *scrutinizing* data (54, 20%). As explained in section 5.1.5, *scrutinizing* denotes the second look when users rigorously scrutinize the utility of the identified dataset, including examining the file formats, data types, and data accuracy; assessing the variables that can combine multiple datasets.

By analyzing the forum posts in-depth (shown in figure 32), we found the top two challenges under *scrutinizing* process were when users identified the sought-after topic of data, they found that some data were either missing (21, 38%) or the time was inaccurate or unwanted (21, 38%). E.g., *Hello, I am having trouble downloading the shooting dataset. It is only downloading a set of 88 records (see attached). There is no filter on that I am able to manage, but one seems to be applied (see screenshot). I've tried this on a few machines and browsers and got the same thing.* The following response confirmed this issue and highlighted that the metadata appeared to contain all records under the "shooting" dataset. However, the user was only able to retrieve part of it, which indicates some data were missing from the file. Additionally, an example of examining the time is that *we are looking to update our traffic victims web page using the Police Fatal Crash data set. The description indicates that it was supposed to be refreshed nightly, but the data has not been updated since it was added on September 25th.* This example shows that the user inspected the expected update time of the dataset and found the portal failed to update the data in a timely fashion.

The third asked questions were about the specific content of the data (8, 14%). E.g., *"Does anyone know if there a possibility that older crime incidents are now being added to the dataset? After downloading the initial CSV file, I am using the API to only get the most recent 3 days worth of data using the dispatch_date_time value. Since there is no "Update Date" field in the dataset, how do I know if some older data was added/changed retroactively?"* At last, users also had concerns about data accuracy (6, 11%). E.g., *"it looks like the Shooting Victims dataset contains duplicates. There should be about 8,500 victims to date, but there are 17,000. Sorting by event date and time clearly shows duplication. Does anyone know when this might be cleaned up?"*

It is noteworthy that all the sub-challenges in *scrutinizing* process manifested that a requisite description for the dataset is essential. By using the example mentioned above, the user asked that *"Since there is no 'Update Date' field in the dataset, how do I know if some older data was added/changed retroactively?"* We can see the importance of the dataset description. In addition, we admit that data is not perfect, but OGD portals should work to provide datasets without mistakes, such as duplicated values and the inconsistency between metadata description and data.

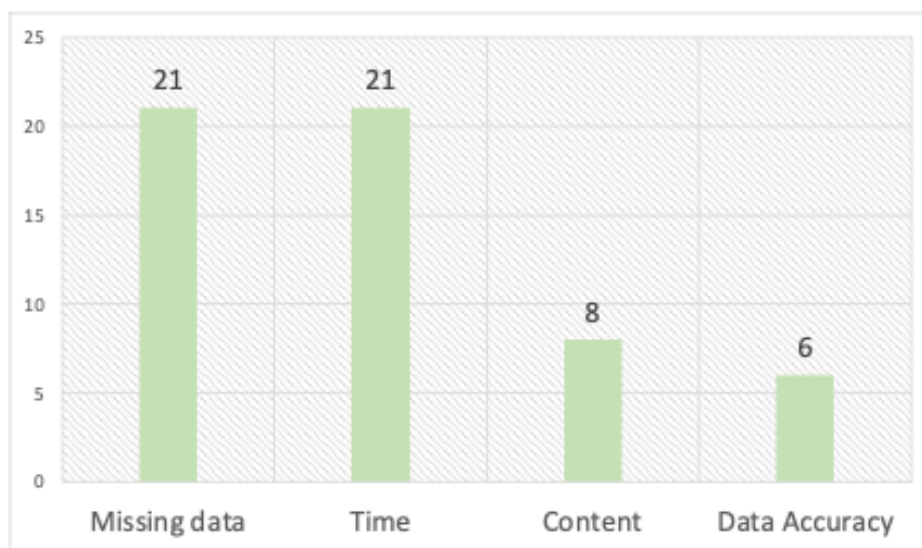


Figure 32: The Sub-categories of *Scrutinize* Challenges

Interview data analysis displays several of the main challenges of *scrutinizing* data. The first one is examining data files. For example, *"the other challenge was because I was so new to coding, I was really unsure what file types to look for, so I wasn't totally sure if CSV would work or if I needed something else."* (P7). Also, for the challenges of assessing the variables that can combine multiple datasets, P4 depicted that *"I think the biggest challenge is trying to find the common denominator in both datasets, so what correlated and the title of ours was, like WDM, and there was something, again, completely different, but it was the same numbers and the exact same data."* P7 has the same barrier of ensuring the collected datasets are all compatible. Group 3 (P8, P9, P10) mentioned they needed geographic information for their project, but WPRDC did not provide this file format, so they had to spend lots of time

exploring the other sources. P17 came up with the barrier of lacking detailed information, e.g., *“for something a little detail, it is not easy to find. Even the data I did find, I found it very insufficient. It didn’t contain critical information.”*

Based on the descriptions of the interview data, we found that in addition to providing correct data and the requisite description of data from the portal side (revealed by forum posts analysis), some OGD literacy skills are desired when scrutinizing data, such as knowledge of data file formats, data types, and the conditions of merging data.

5.2.3 The Challenges in the Context of *Data Sensemaking Stage*

5.2.3.1 The Challenges in the Context of the *Understanding Process*. The top three challenges of interacting with OGD are all in the *data foraging stage*, which echoes prior studies that “data is hard to find” (Lammerhirt, 2017; Zuiderwijk, Janssen, et al., 2012). The fourth identified barrier is *understanding* data (29, 11%) which is in the *data sensemaking stage*. *Understanding* denotes the process when users start to perceive the intended meaning of the data, such as understanding the meaning of the column headings and the values of the data.

By deeply breaking down the related posts, we detected six sub-categories of user barriers to understanding data (shown in figure 33). The result displays that the meaning or scope of the values is confusing to users (15, 47%). E.g., *“anyone knows what these values represent? They’re integers -1 through 9, there’s no description on the metadata.”* Or *“Does anyone know what the following case codes mean for LI? ABATE, ACTC, ...”*

Similar to the values of data, the second asked question is about the meaning of column headings or field names (9, 28%). E.g., *“Can someone please define what the field names are for this dataset? Specifically, I’d like to know what PCPC_Num stands for, but it would also be helpful to know the rest.”* We further examined the posts related to the values and column heading, including the initial questions and responses, to explore why the users encountered the challenges of understanding them. We found that in most situations, users knew to check the metadata information, but the portal websites failed to provide enough information to assist users in understanding the data, which was discussed extensively in previous studies

(Eberhardt & Silveira, 2018; Gruen et al., 2014).

The forum data analysis also found that users sought the description of the data collection process (5, 16%) to help themselves better understand or trust data. E.g., *“Is this a subset of data from the state dataset? If it is, can you explain how you are collecting the data, and which methods are being used to subset? If it’s data directly from Philadelphia, can you confirm that the records match the state records?”* As presented in figure 33, we also observed that users encountered the challenges of understanding the provided description (1, 3%) and visualizations (1, 3%), and one user was curious about how the data is stored in the database (1, 3%).

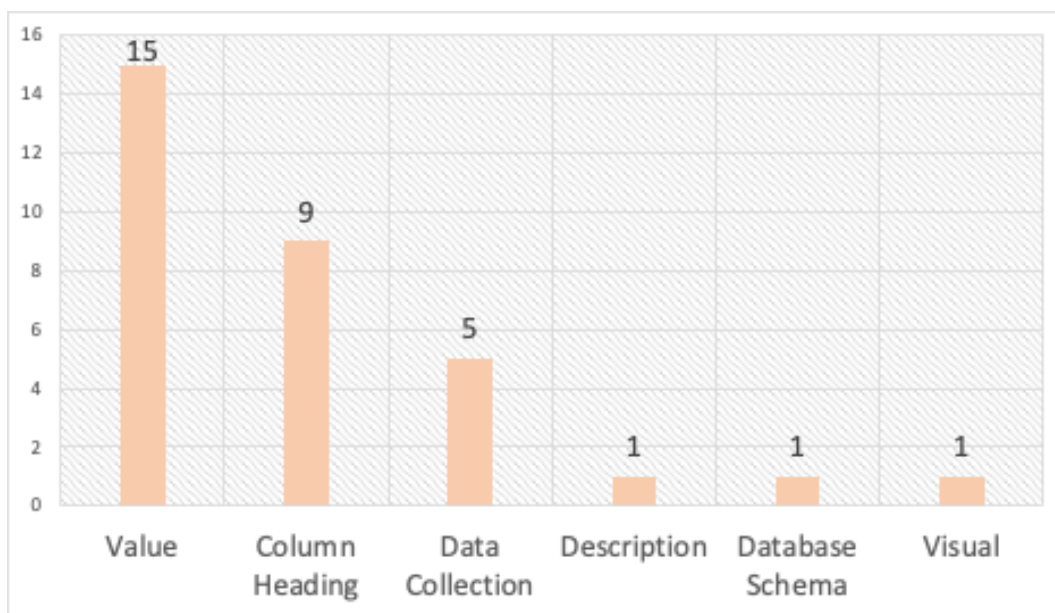


Figure 33: The Sub-categories of *Understand* Challenges

The interview data analysis also found that the major challenge for participants is understanding the meaning of column headings and their values, primarily because of the insufficient information that data portals provide. For example, *we did have a couple of different ways of getting around; since WPRDC doesn’t have all the information we wanted, we did go to the county website and looked for information (group 3: P3, P4, and P5)*. Another critical challenge was pointed out is the lack of transparency for some events (P13). For example, *“the data in the file doesn’t match the file name, so in the event that happened,*

I had to figure that they didn't get new data today or other assumptions... So, I wish they could give a statement to say hey, if, in the event that happens, this may be the reasons why."

This participant complained that it happens very often. He explained that he understands that some data need to be anonymous, and sometimes, even cannot be given an apparent reason for the anonymity or just giving a null value. However, a general statement from OGD portals that informs users that this is not an error would be helpful.

By synthesizing the results from forum posts analysis and interview studies, to address the challenges in the *understanding* process, providing sufficient information about data by OGD platforms are essential.

5.2.3.2 The Challenges in the Context of the *Schematizing Process*. Another interactive challenge in the stage of *data sensemaking* is *schematizing* data. Based on the new H-OGD-I model, *schematizing* data indicates the process when users try to establish the logical relationships between the data, namely, organizing the data with relevant information or building connections between the data.

Forum data analysis shows that it is hard for users to build connections between data (11, 4%). E.g., *"I noticed there is no column for the total assessed value for each property. I was wondering if I could extrapolate that information by adding up taxable land and taxable building or if I should reference another field."* This type of question could be addressed by providing a comprehensive context information document to elaborate more detailed information about the data. On the other hand, another type of question related to *schematizing* data may require user literacy skills to solve. E.g., *"I see polygons- labeled Loss, Gain, or No Change- but I'm wondering how those characteristics affect the area of the polygons. For example, can I assume that a polygon labeled Loss was larger back in 2008?"*

Interview data analysis proved the finding that *schematizing* data is challenging. For example, P3 described that *"interpreting what I am looking at and what that (data) means is challenging... Once you have all the information side by side, looking at all the different numbers and changes, wondering, what do these mean?"* Also, the difficulty can be generated simply because the data size is too big. E.g., P4 complained that *"there were upon thousands of columns and fields in the polling place data; even just looking at the column headings*

requires a lot of scrolling, and searching, so the initial analysis aspect was a challenge because we had to pull out what is relevant, a lot of stuff that was in the list of a polling place was...

In general, schematizing relies more on user OGD literacy skills.

In a nutshell, in the stage of *sensemaking*, the portal side (services, design, functions) plays a more critical role in the process of understanding data, while the user side (OGD literacy) is more essential in the process of schematizing data.

5.2.4 The Challenges in the Context of the *Data Use Stage*

5.2.4.1 The Challenges in the Context of the *Preparing Process*. Forum posts analysis found that *preparing data* (13, 5%) is the most challenging barrier in the stage of *data use*. After further analysis, we identified six different granular challenges in the *preparing* process (presented in figure 34). Importing GIS Data or shapefiles to a platform such as ArcGIS or other map tools was the top-asked question when preparing data. E.g., “*I’ve tried downloading a number of shp files for data from the City of Philadelphia that will not import to ArcGIS Pro through feature class to feature class.*” Also, users sought help to merge (3, 23%), extract (2, 15%), and clean (1, 8%) data. One user (15%) failed to transform data from a CSV file to a shapefile as the data missed the zip code values.

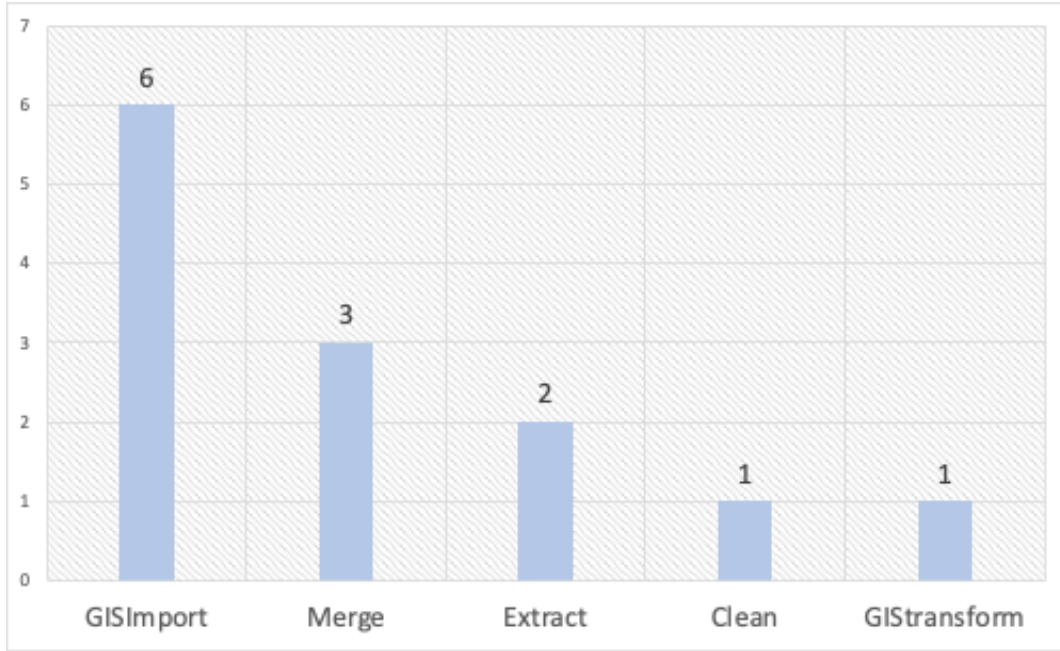


Figure 34: The Sub-categories of *Prepare* Challenges

The interview data analysis revealed similar findings. The difficulty of extracting data was identified. For example, P7 depicted “... *that would be really the only challenge that it was just deciding which parts of the dataset I needed and how to extract it out.*” Also, P4 and P5 described that “*I think our biggest challenge is merging the datasets, we tried numerous different methods. Pretty much each of us was working and trying to find a way through Python.*” In addition, cleaning data was also proposed, “*I would say there was one point where I had to add those last two zeros to a bunch of data points... at the very end, I spent so much time trying to figure out that*” (P1). All these challenges point to the need of advancing users’ OGD literacy skills.

5.2.4.2 The Challenges in the Context of the *Analyzing Process, Representing Results Process, and Communicating Process.* The other two challenges disclosed from forum data analysis in the *data use* stage are analyzing data (6, 2%) and representing results (3, 1%). Furthermore, the challenges of communicating data and using programming languages were identified by interview data analysis.

Most of the posts regarding analyzing data were describing their data analysis process, and then asking a related question. E.g., *“I am not understanding how to correlate the location data... I have the census_block_group as a column in the data set but that’s about it. I also read about Geo-IDS. I am now confused about how to proceed. Any inputs would be highly appreciated.”* In addition, two questions about representing results were asking how to make visualizations. E.g., *“From response to a previous post, I was using PA83-SF. On close visual inspection in AutoCAD Civil 3D 2022, the footprint polylines are not at a 90 degree angle (see picture). Please help me understand what I am doing wrong or provide me with the correct Global Coordinate System Code.”* Someone also asked for support on how to describe data. E.g., *“Thank you for the reply and pointer on the threshold of 6. Do you have any advice on how to report on these low numbers that default to 6?”*

Similarly, analyzing data was laid out in interview data analysis. For example, *“The first thing I did was to google how to make run a correlation; that is an easy way to see if it matches. But then, we did have to figure out other ways to compare the data, which is the biggest challenge.”* (P2). Also, appropriately representing results was difficult; P4 and P5 stated that *“using the vocabulary to describe what we were seeing and what we were talking about was difficult. Just because we wanted to be politically accurate, and we wanted not to say anything bad.”* The challenge of communicating data was pointed out by P15, *“I think the other challenge would be telling a story beyond what was represented on the data portal already.”* Using programming languages to processing data is also a big challenge for the participants. All three group participants (all students) proposed using Python is hard, which makes using data difficult. For example, *“the biggest challenge was using Pandas...”*(P1).

On the other hand, we did have four participants (P14, P16, P17, P18) who claimed that they might have challenges with other tasks, but regarding the exemplified task for this interview, they did not encounter any difficulties in using data. After reviewing their familiarity with OGD, all these four participants have used OGD over 15 times. Although familiarity with OGD cannot represent OGD literacy, this observation still shows that with more experience using OGD, users have fewer challenges in using data.

Overall, as presented in figure 30, challenges encountered in the stage of *data use* were

brought up less than in other behavioral stages on the forum. However, this observation cannot prove that users have fewer questions in using data in practice. On the contrary, they could encounter many challenges in this stage. Our interview data analysis revealed more difficulties in the *data use* stage. During the interviews, we could feel participants' intense frustration or pressure when describing the challenges of interacting with data. Therefore, we drew the conclusion that users have more challenges in the *data use* stage than what the forum data accounts for. The reason for fewer questions on the forum posts could be because users may not think an online forum is the right place to ask complex skill questions, as using data requires specific user skills, which might be addressed better elsewhere.

After examining all the contextualized challenges, we found that these interactive challenges can be overcome through three primary high-level aspects: data policy, platform interface, and function design, and OGD literacy skills. Some challenges may only be solved by data policy, such as releasing more data. Some challenges more rely on the design or information provided by OGD portals, such as visualization tools and metadata information. Whereas, other challenges may more depend on user OGD literacy skills. For example, knowing basic data knowledge and performing mathematical statistics. Since this dissertation is to investigate OGD users from a user-centered perspective, I explored and identified the foundational OGD literacy capabilities in Study 5, and the results shown in RQ3 (5.3).

5.2.5 A Summary for RQ2

To clarify, the sum of sub-challenges in each behavioral process might be higher than the frequency for each process, because one post could cover multiple sub-challenges. For example, the frequency of *understanding* challenge is 29 (see Figure 30), but the sum of sub-challenges in the process of understanding data is 32. One post could indicate the challenges of understanding the meaning of values and column headings. Furthermore, the total number of coded posts (269) is higher than the initial post number (238). That is because one post could contain multiple behavioral processes and corresponding challenges. For example, with this post – “*the OPA datasets do not have an OPA Account Number column. The only unique id columns for these datasets seem to be "Registry Number" and "ObjectID". Where can I*

find information on the OPA Account Number for this dataset”, we classified it as *Find* and *Scrutinize*. Therefore, this post was counted twice.

After examining all the contextualized challenges, we found that these interactive challenges can be overcome through three primary high-level critical facets: data policy, platform interface and function design, and OGD literacy skills. Some challenges may only be solved by data policy, such as releasing more data. Some challenges rely more on the design or information provided by OGD portals, such as visualization tools and metadata information, whereas other challenges may more depend on user OGD literacy skills, such as knowing basic data knowledge and performing mathematical statistics. As this dissertation is to investigate OGD users from a user-centered perspective, I explored and identified the foundational OGD literacy capabilities in Study 5, and the results are shown in RQ3.

5.3 What are the Fundamental OGD Literacy Capabilities that Enable Users to Use OGD? (RQ3)

To explore and identify the needed fundamental OGD literacy capabilities, we conducted a content analysis on the online forum posts and the interview data. Of the 415 forum posts qualifying for analysis, most of them were the responses. That is because we found that responses generally include the solutions to the questions, which revealed the required capabilities that help overcome the difficulties. We also examined the initial questions posts and included them in the analysis, as long as the posts contained descriptions indicating users’ abilities. Figure 35 presents the distribution of the OGD literacy capability in each behavioral process; in descending order, they are the processes of *finding* (151, 28.8%), *scrutinizing* (100, 19.1%), *acquiring* (74, 14.1%), *developing data needs* (72, 13.7%), *understanding* (54, 10.3%), *using* (50, 9.5%), *schematizing* (14, 2.7%), and *sharing data* (9, 1.7%). Because one post can manifest multiple capabilities, the sum of the capabilities identified in the H-OGD-I processes (524) is higher than the initially identified post number (415).

In the interview instrument, the questions associated with OGD literacy capabilities were designed to further elicit the required user OGD literacy capabilities. Additionally, the

OGD literacy capabilities needed in the processes of *exploring existing datasets* and *framing data-centric questions* that could not be identified by forum data analysis were revealed by interviewees. After synthesizing the results from the forum data analysis and interview studies, we organized two focus groups with experts in data literacy and OGD to evaluate the findings.

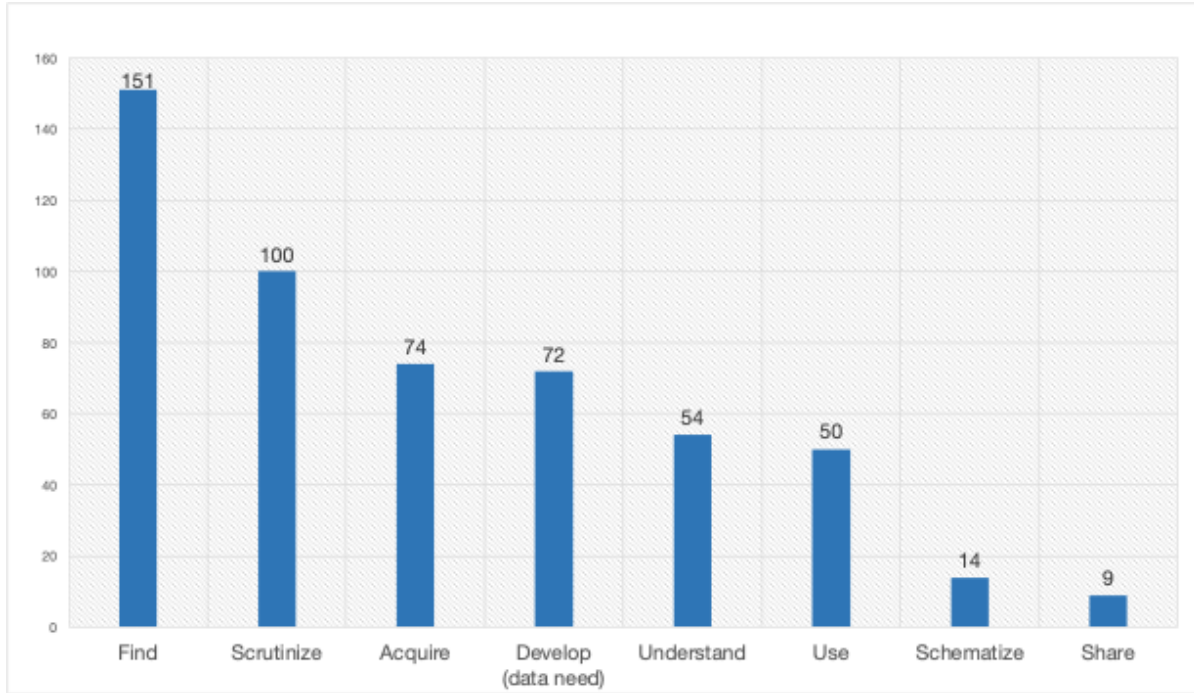


Figure 35: The Distribution of the OGD Literacy Capability in Each Behavioral Process

In addition to the capabilities identified in each H-OGD-I behavioral process, defined as *Action-specific OGD Literacy Capabilities* (code: ASC-N), we also discovered essential abilities or mindset required throughout the entire process or multi-processes, defined as *Action-holistic OGD Literacy Competencies* (code: AHC-N). Three categories of Action-holistic OGD Literacy Competencies are **Conceptual Foundations, Advanced Actions, and Dispositions**.

This section describes the results of the *Action-specific OGD Literacy Capabilities* and *Action-holistic OGD Literacy Competencies*, based on the two empirical studies and the suggestions from the two focus groups.

5.3.1 The Fundamental Action-specific OGD Literacy Capabilities in the *Task Preparation Stage*

5.3.1.1 The Capabilities in the Process of *Exploring Existing Datasets*. Due to the nature of the online forum, the behavioral process of *exploring existing datasets* in the *task preparation* stage cannot be manifested or inferred from the forum posts. As explained previously, users asked more questions about data per se on the online forum rather than at the very beginning of preparing for the task. Therefore, the findings were primarily uncovered from interview data analysis.

Based on the new H-OGD-I model, *exploring existing datasets* in the *task preparation* stage denotes the process when users intend to propose a real problem that interests them by using the existing datasets. The typical behaviors include browsing the categories or tags of datasets and evaluating the topic's relevance. We considered that when a user can carry out those typical behaviors and finally achieve the goal, we would acknowledge those abilities as OGD literacy capabilities. For example, P3 described that “*we discussed, and we knew that is public information, we know where is polling space is, and we found it at the WPRDC*”. P1 stated that “*I just went to WPRDC, and they have the left-hand columns that describe what types of data they have.*”

In addition, seven data-inspired tasks that refer to the tasks starting from exploring the existing data were revealed in Study 3. For example, “*we looked at the WPRDC, we just kind of saw what was available, I think we formed our basic questions from looking at the polling places data at WPRDC (P5)*”. Therefore, initiating data-inspired tasks indicates that users were able to propose meaningful topics based on the available datasets. Based on the descriptions of participants, we concluded that the fundamental capabilities for exploring existing datasets include 1) awareness of the existing data sources, 2) ability to follow the infrastructure of the datasets, e.g., the classification structure, 3) ability to browse dataset subjects/categories provided by the data portal to obtain inspiration for initiating a project.

5.3.1.2 The Capabilities in the Process of *Framing Data-centric Questions*

For the same reason, the behavioral process of *framing data-centric questions* was not able

to be identified based on the forum posts. Therefore, the findings regarding the capabilities were also mainly discovered through interview data analysis. The three group participants showed me their task reports, which contained a clear list of questions that needed to be answered when performing the tasks. Group 2 represented the following challenge: *“I think the majority of our challenges were when we were trying to figure out our questions, we got all these data, we all thought it was going so well, and then we had the meeting with professors, and they were like it doesn’t really make sense, so then we had to scramble.”* This example signified that to be able to frame data-centric questions is a critical ability when conducting a data task. In addition, in the first focus group that aims to evaluate our findings, the experts stressed the significance of knowing how to ask the right questions. One expert explained that *“a good example would be that students are interested in studying gentrification, so they want to look at gentrification. It is a concept, but then they need to translate that into what are the data that they are going to use, being able to understand the task, and translating that higher-level thing down into something that is a little bit more granular. Getting that point is a major literacy point.”* Eventually, according to participant descriptions and expert suggestions in the focus group, we proposed three fundamental capabilities in the process of *framing data-centric questions*, which are 1) ability to translate an information-based problem into data-centric questions, 2) ability to break a complex problem down into smaller, more manageable questions, 3) ability to accurately identify the right questions.

5.3.1.3 The Capabilities in the Process of *Developing Data Needs*. The behavior of *developing data needs* was inferred by forum posts analysis. According to the coding schema, when the post shows the ability to determine what data is needed to complete the task, we coded this post as “Need.” Forum data analysis identified 72 (13.7%) posts related to the ability to develop data needs that were classified in the task preparation. The following posts indicate the users knew the data need. E.g., *“I am working on a research project and need to find the information (CSV file or otherwise) that contains the full OPA data dump around the second quarterish 2016 to the end of the year 2016. It would likely contain the newly established figures for the 2017 tax year.”* This description clearly states the requirements of the needed data, including the data topic, format, and specific time. E.g.,

“I work for a XXX in South Philly, and we’re working on getting a list of all residential addresses within a given planning boundary to do a neighborhood survey. We’d need to have not only the addresses of single family homes but also units within multi-unit buildings. Does OPA keep a file like this?” This post emphasizes the granular data demands.

The interview data analysis validated the same finding of the forum data analysis for developing data needs capability. The task reports from the three student groups also displayed a clear list of needed datasets with data sources. Moreover, all participants depicted the process of *developing data needs*. Many participants reported specific needs: for example, *I need the data with shapefile format or we needed two datasets with same zip code*. The needs include data content, types, and time, to name a few. Interview participants also explained the essentials of knowing data needs. For example, P7 stated that *“I think certainly having an understanding of what you need might be the first step. It was really like defining to yourself what you need and in the simplest terms that could then be used to search for the data that you need.”* P11 considered that *“be able to identify what you have and what you don’t have and what you need” is a critical OGD literacy at the beginning of performing a task.* Given those descriptions about forming data needs, we concluded the capability in the process of developing data needs as ability to determine what data is needed to complete the task.

To clarify, the behavior of *developing data needs* was counted under *finding* behavior and challenge when annotating the forum posts in Study 2 and Study 4. That is because when users posted the description about data needs, they were seeking help in retrieving data, and the “data needs” description was to explain what data they were looking for. However, in analyzing capabilities, *developing data needs* and *finding data* were broken down into two separate abilities, as figuring out the required data needs is a critical capability to perform a task. Table 8 presents the fundamental OGD literacy capabilities in the stage of *task preparation*.

Table 8: OGD Literacy Capabilities in the Stage of *Task Preparation*

Behavioral process	Actions	OGD literacy capabilities	Codes
Explore existing datasets	Browse Datasets	Awareness of the existing data sources	ASC-1
Explore existing datasets	Browse Datasets	Ability to follow the infrastructure of the datasets, e.g., the classification structure	ASC-2
Explore existing datasets	Browse Datasets	Ability to browse dataset subjects/categories provided by the data portal to obtain inspiration for initiating a project	ASC-3
Frame data-centric questions	Translate a real problem to data-centric questions	Ability to translate an information-based problem into data-centric questions	ASC-4
Frame data-centric questions	Translate a real problem to data-centric questions	Ability to break a complex problem down into smaller, more manageable questions	ASC-5
Frame data-centric questions	Translate a real problem to data-centric questions	Ability to accurately identify the right questions	ASC-6
Develop data needs	Frame Data Needs	Ability to determine what data is needed to complete the task	ASC-7

5.3.2 The Fundamental Action-specific OGD Literacy Capabilities in the *Data Foraging* Stage

This section argues the needed OGD literacy capabilities in the stage of *data foraging*, which contains four behavioral processes: *find*, *evaluate*, *scrutinize*, and *acquire*. We created a code for evaluating data in the coding schema, that is Eva-Topic, referring to the ability to evaluate if the identified data is topic relevant. However, our forum posts and interview data did not identify the challenges and capabilities in this *evaluating* behavioral process. After discussions with collaborators as for the possible reasons for no evaluation-related capabilities discovered, we concluded that evaluating the topic-relevance of data requires

more reading comprehension skill. Therefore, this section illustrates the capabilities required in the processes of *find*, *scrutinize*, and *acquire*.

5.3.2.1 The Capabilities in the Process of *Finding* Data The forum data analysis demonstrates that the capabilities in finding data process were discussed the most in the online forum (151, 28.8%, shown in figure 36). Based on the coding schema (section 4.7.2), we explored the forum posts and identified 151 posts that can infer the associated capabilities. Specifically, 115 posts showed the capability associated with Find-Source-ES, which refers to users being aware of the existence of data and/or knowing where to find the needed data. The following posts indicate the users knew the needed data existed and had already located them. E.g., *“I was wondering if you could tell me a little more about variables in the Philadelphia Street poles data located here: <https://www.opendataphilly.org/dataset/street-poles>.”* This post obviously manifests this user already identified the sought-after dataset. E.g., *“Is there an estimate for when 2019 salary data will be uploaded?”* This post displays that the user was aware that the salary data exists and was just trying to find the particular time.

In addition, the forum posts discovered 33 posts that can infer the capabilities associated with Find-Source-MUL, which refers to users being able to identify and locate multiple datasets from the same or different sources to fulfill the data need. E.g., *“I think you can find this in the Parks and Recreation Assets data set at <https://www.opendataphilly.org/dataset/parks-and-recreation-assets>. You can find other parks and recreation data sets at <https://www.opendataphilly.org/dataset?tags=Philadelphia+Parks+and+Recreation>.”* This post denotes that the user knew where to find the needed data in different datasets.

We also observed one post proposing browsing data (Find-Seek-Browse), two posts mentioning finding data through advanced search (Find-Seek-AS), and one post displaying searching data by using queries (Find-Query-KWs). This observation shows that only a few posts on the forum describe the seeking methods. However, according to Study 1, browsing and keyword searching are the most used seeking behaviors. Therefore, this could be a limitation of online post data. More insights of the needed capabilities for *finding* data were revealed through interview data analysis.

Interview data analysis validated the findings from the forum data analysis and also compensated its limitation by directly clarifying the required capabilities for seeking data. Regarding awareness of the existing data and data sources, P2 noted that *“It is helpful that even if you haven’t done a data analysis project, just knowing that certain government data, and you can go and find it.”* P10 further stressed the reliability of data sources, she expressed that *“it is important to understand the reliable sources for data.”* Similarly, P14 stated that *“Knowing where to get data that is standardized across a large geographic area can smooth out mistakes.”* Interview data analysis also presented the needed skills in online browsing and searching. For example, for online browsing, P6 claimed that *being fairly strategic. If I looked at a dataset to consider it for use and it seemed to have a problem, I just rejected it and moved on. By the time the process was complete, we had arrived at those datasets that were solid and good, at least for our needs.* For online searching, P9 described that *“I think having the correct or accurate keywords to look for data is a difficult start. Because if you have no idea what you are going to search for or what keyword you are going to enter, it will be very hard to get useful and reliable data.”* P4 claimed that *“we had to find the data for our project, various searches and everything, so I think finding the data itself is definitely can play a role in data literacy aspect.”*

Based upon the forum data analysis and interview data analysis, we concluded six fundamental capabilities in the *finding* data process: 1) awareness of the existence of data and/or knows where to find the needed data, 2) ability to identify and locate multiple datasets from different sources to meet the data needs, 3) ability to choose keywords or various combinations of keywords to find the needed data, 4) ability to conceptualize the infrastructure of the datasets, e.g., the classification structure, 5) ability to leverage the provided categories/groups/tags to look for data (knows data categories), and 6) ability to combine search methods (e.g., keyword search + subject browse or keyword search + cited-reference search) to seek data.

5.3.2.2 The Capabilities in the Process of *Scrutinizing* Data. Forum data analysis identified 100 posts (19.1%) indicating the capability of *scrutinizing* data. According to the coding schema (section 4.7.2), SCR-Utility refers to the user being able to rigorously

scrutinize the utility of the identified data, including identifying file formats, data types and data time period; examining if important data values are missing; checking the accuracy and comprehensiveness of the data; and assessing the variables that be used to join multiple datasets. The following posts indicate the users are able to examine data in these various aspects. E.g., *“The default download from there doesn’t have the latitude and longitude that the csv download from the city provides, however, which has been useful in creating links to the Pictometry data.”* This post manifests that this user inspected the missing data from the downloaded dataset. E.g., *“I’ve noticed the latest date in the full dataset is Dec. 31, 2017. And the files titled ”2018-present” don’t have any entries.”* This post indicates that this user was able to examine the time and missing values of the located dataset. E.g., *It looks like the Shooting Victims dataset contains duplicates. There should be about 8,500 victims to date, but there are 17,000. Sorting by event date and time clearly shows duplication.* This post infers that the user had the ability to probe into the data accuracy. We also observed that users were able to review the description of datasets, and determined the accuracy of the data time. E.g., *“the description indicates that it was supposed to be refreshed nightly but the data has not been updated since it was added on September 25th.”*

Interview data analysis confirmed the findings from forum data analysis. Several participants brought up the ability to scrutinizing data formats, types, and size. For example, P2 proposed that *“being able to know what file types are, like shapefile, this is for mapping.”* Similarly, P3 came up with the need to have *“some types of knowledge of file formats, csv or xml.”* In addition, the capability of assessing the variables that can be used to combine multiple datasets was discussed by participants. For example, P3 stated that *“we needed to know we can merge those datasets, and all of the columns would match up, and all of the rows would match up. So, without that knowledge, I wouldn’t know to merge and compare the information.”*

Based on the analysis of the forum data and interview transcripts, we found four capabilities in the process of *scrutinizing* data: 1) ability to scrutinize if the data file formats, data types, and data time period satisfy the needs of the task, 2) ability to determine if data fields or values are missing, 3) ability to check the accuracy of the data, and 4) ability to identify and assess fields that can be used to join multiple datasets (used in the “Merge”

capability).

5.3.2.3 The Capabilities in the Process of *Acquiring* Data. Forum data analysis revealed the needed capabilities of acquiring data and identified 74 posts (14.1%) that display the pertinent abilities, specifically, manually acquiring data by leveraging the provided links or following instructions (48 out of 74), and obtaining data by utilizing programming languages, e.g., API, SQL (29 out of 74). Some posts consist of the combination of the two main methods to acquire data, thereby, the sum of $48 + 29$ is higher than the initially identified posts (74). The following post reflects that the user can manually acquire data:

It looks like it keeps showing the data in a table in HTML. Here's I got around it:
1) go to <https://github.com/CodeForPhilly/RAPper/blob/master/data.csv> (which is linked from opendataphilly)
2) highlight the html table and copy the data
3) paste it into your favorite spreadsheet (I used Google Sheets)
4) Fix the first row so it lines up with the rest of the table and delete the blank first column
5) save file as CSV and then it's good to use

In addition, being able to acquire data by using programming languages is a key capability to obtain a large size of data. The following post manifests the user can acquire data by using API:

If you're trying to download through browser, you may be hitting a timeout due to the database size.
If you are comfortable working with Carto's SQL API, you may have better success via that route: <https://carto.com/developers/sql-api/>
If you can't get it working with a simple POST query, maybe look at the Batch Query API <https://carto.com/developers/sql-api/reference/tag/Batch-Queries>

It is noteworthy that based on interview data analysis, none of the participants actively proposed the capabilities for acquiring data. Most of the participants simply clicked the downloading links to obtain data without difficulties. Some participants thought if the downloading link did not work, it is the platform issue rather than the lack of the data literacy skills. There were several participants using APIs to acquire data, such as P1, P8, P14. Even though, P1 did mention that “*I didn't choose based on the data itself, I chose based on how easy the API was to use, and how efficient it was,*” she did not define this as a data literacy skills. However, one expert in the focus group who is also working for

a data center considered that using API to acquire data is an essential data literacy skill. Ultimately, we concluded two capabilities for acquiring OGD: 1) ability to download (e.g., through provided links) or collect the needed data manually, and 2) ability to acquire the needed data by using a programming language (e.g. API, SQL).

In total, 12 capabilities were developed in the stage of *data foraging*, shown in Table 9.

Table 9: OGD Literacy Capabilities in the Stage of *Data Foraging*

Behavioral process	Actions	OGD literacy capabilities	Codes
Find	Find Sources	Awareness of the existence of data and/or knows where to find the needed data	ASC-8
Find	Find Sources	Ability to identify and locate multiple datasets from different sources to meet the data needs	ASC-9
Find	Search with Keywords	Ability to choose keywords or various combinations of keywords to find the needed data	ASC-10
Find	Browse Data Subjects	Ability to follow conceptualize the infrastructure of the datasets, e.g., the classification structure	ASC-11
Find	Browse Data Subjects	Ability to leverage the provided categories/groups/tags to look for data (knows data categories)	ASC-12
Find	Combine Search Techniques	Ability to combine search methods (e.g., keyword search + subject browse or keyword search + cited-reference search) to seek data	ASC-13
Scrutinize	Scrutinize Usability	Ability to scrutinize if available file formats, data types, and data time periods match the needs of the task	ASC-14
Scrutinize	Scrutinize Usability	Ability to determine if data fields or values are missing	ASC-15
Scrutinize	Scrutinize Usability	Ability to check the accuracy of the data	ASC-16
Scrutinize	Scrutinize Usability	Ability to identify and assess fields that can be used to combine multiple datasets (used in the “Merge” capability)	ASC-17
Acquire	Acquire Data	Ability to download (e.g., through provided links) or collect the needed data manually	ASC-18
Acquire	Acquire Data	Ability to acquire the needed data by using a programming language (e.g. API, SQL)	ASC-19

5.3.3 The Fundamental Action-specific OGD Literacy Capabilities in the *Data Sensemaking* Stage

5.3.3.1 The Capabilities in the Process of *Understanding* Data. Forum data analysis pinpointed 54 posts (10.3%) that can show the required OGD literacy capabilities in the process of *understanding* data. After further examination, three primary abilities emerged. Based on the coding schema (section 4.7.2), the first key ability is SenseMU-Context-4w1h (34 out of 54), referring to a user being able to know when, where, and how the identified dataset was collected, who collected it, and what the dataset is about, the scope of the data, and being able to comprehend the context information, and use it to help understand the data. For instance, *“I can’t speak to the ordering of rows on the server. I know that CARTO uses managed cloud services, and they may not even have control. My guess is that the police department is simply appending new rows each day, so it may regency the way I’m which the historical data was originally loaded.”* This post manifests that this user had some background knowledge and could use it to help understand the context of the dataset.

The second critical skill is SenseMU-FieldNV-Dictionary, which indicates that a user can understand the meaning of column headings and/or the corresponding values by using provided data dictionary or metadata (11 out of 54). For instance, *“The metadata for the 1, 2, and 3-digit codes is in the metadata record for this data set accessible from the OpenDataPhilly web page you listed.”* This post shows that this user knew that the meaning of column headings and values should be found in the metadata or data dictionary and knew to seek answers through those files.

However, the metadata or data dictionary on the website could be incomplete. In fact, most participants thought that the data dictionaries offered by OGD portals were not comprehensive or did not fully address their needs, which caused users to be unable to understand the data. In this situation, the third essential capability plays an important role. SenseMU-FieldNV-Personal denotes that a user is able to understand the meaning of column headings and/or the corresponding values based on personal abilities, such as previous background knowledge, comparing with a similar dataset, and Google search (15 out of 54). For in-

stance, *“Those are fields that are automatically generated by ArcGIS for polygon shapefiles. Shape_area is area (probably in square feet) and Shape_length is perimeter (probably in feet).”* After examining the initial question and all the responses, we determined that this user answered the question based on his background knowledge.

The posts with sub-coded capabilities overlap with each other. Therefore, the total sum of the sub-codes (60) is higher than the identified post number (54).

Interview data analysis also discovered the capabilities associated with being able to understand the context of the data and the meaning of the column headings and values. For example, P7 considered that *“it is essential to understand the reason that data was created and the people that collected it, why they were collecting it, and what is the original purpose of the data.”* When discussing the critical OGD literacy capacity for completing the task, P13 claimed that *“I think it is impossible to know this unless you have a background in public health.”* These two examples manifest the requirement for domain background knowledge to better understand data. Participants also proposed the capacity of understanding the meaning of column headings and values. For example, P8 reported that *“understanding the importance of metadata and the value of that read-me file, like all data explanation is critical.”* P3 thought OGD literacy involves *“some knowledge with a data dictionary.”* In the case when the data dictionary failed to fully define or clarify what the column headings mean, P15 told us an approach to address it. That is *“looking at comparable datasets that were shared in other data portals to see if perhaps the data dictionary is available on those portals that might provide a lens to the data that you were looking at, so looking at similar data”.* This post indicates that professional users discover various ways to understand the identified data. On the other hand, P14 claimed that *“for understanding data, my view is that it is the responsibility should be all on the publisher. For example, a publisher in its metadata writes something that is not clear that would not be clear to someone from the public; they should make sure to clearly explain the topic in the metadata.”* According to all those ideas, we concluded that the publishers need to provide transparent and comprehensive information for describing datasets, and users need to know the existence of the data description documents and their usage.

In addition, the perspective of a certain level of comfort with technology (P7, P14) and/or

statistical skills (P9, P13) can be advantageous to understanding data was brought up by participants. For instance, “*if you have statistic literacy, at least you can read data, like quantitative data*” (P9). P14 stated that “*knowledge of spatial and non-spatial data, such as knowing the spatial nature and non-spatial, can make a big difference.*” Based on the findings from forum data analysis and interview data analysis, we summarized three fundamental capabilities for the *understanding* process: 1) ability to understand when, where, and how the identified datasets were collected, who collected them, and what the dataset is about, 2) ability to comprehend this context information, and use it to help understand the data, 3) ability to understand the meaning of column headings and/or the corresponding values by using a provided data dictionary or metadata, or based on personal abilities, such as previous background knowledge, and Google searches.

5.3.3.2 The Capabilities in the Process of *Schematizing* Data. Forum posts analysis discovered 14 (2.7%) posts that showed the capabilities of *schematizing* data. In the coding schema, we designed two sub-codes: SenseMS-Logic (11 out of 14) and SenseMS-Visual (3 out of 14). SenseMS-Logic refers to a user being able to 1) build connections between the columns and/or tables and 2) understand the relationships between data by manipulating data, such as sorting or filtering. The following posts manifest the users were able to make sense of the numbers by comprehending background information and manipulating data. E.g., “*It makes sense now. We know the total from cases by outcome (as of 6/10/20, it was 1,433), and any difference in totals from the stratification files is due to the suppression for counts under 6.*” This post tells us the user was making sense of how the number was calculated. E.g., “*For instance, the 10/29 numbers are 4111 Negative and 447 Positive which by my math would be 4558 Total with a Positivity % of 9.81% (double-checked by percentagecalculator.net) but the “Positivity %” purple line graph for 10/29 lists 7.7%.*” This user did the math and compared the results with the graph on the website.

SenseMS-Visual signifies that a user is able to 1) know the basic visualization categories, such as bar chart, line chart, heat map, and geographic map, 2) read the information on the Scale, Legend, and Axis; and identify the data format that used in the visualization. The following posts indicate the users were able to digest information from visualizations.

E.g., *“I can’t find Booster Vaccination Statistics under COVID-19 Vaccination Dataset, even though I see the booster graphs in the Visualization Dashboards.”* This post shows that this user found some missing data based on understanding the corresponding visualization. E.g., *“The other idea is that you could use the visualization to narrow down the number of records by adding filters, just clicking on the map to select a certain area, or selecting a timeframe.”* This post presents that this user not only understood the visualization but could also use it as a tool to filter the wanted data.

Interview data analysis discerned similar discussions concerning the capabilities of schematizing data. Participants proposed the capability of building logical relationships. For example, P7 described that *“being able to understand the structure of the data, and organize those structures is helpful to make sense of data.”* This participant also pointed out that *“being able to analyze what is in the data in comparison to your needs requires that you be able to think about it in more than what is just represented to you and not by altering the meaning of the data, but by using that to support your own project or needs.”* This participant stressed that understanding the structure of the data would assist in interpreting the data. P18 described that *“be able to compare signal points with the entirety is important. For example, I wanted to know the safety of our neighborhood, I saw the numbers of criminal instances, but the number didn’t make much sense to me until I compared it with other neighborhoods or cities. I would have a sense if our neighborhood is good or not.”* Those discussions indicate that building connections between data are critical to a deeper understanding of what the data really mean.

Participants also discussed the capability of understanding visualizations. For example, P2 expressed that *“I think that people see things like graphs and charts that kind of stuff all the time, everywhere, so knowing what is behind that, I mean a lot of time that stuff is showing what an author intends to show which may or may not be the entire picture.”* This participant put forward a critical argument on how to absorb the information from visualizations. P17 also stressed the ability to read visualizations by depicting that *“being able to look at charts and tables is important. If you don’t understand charts and tables, you are probably not going to ever find what you are looking for.”* By combining the findings from forum data analysis and interview data analysis, six capabilities for the *schematizing* process were developed,

containing 1) ability to build connections between the columns and/or between tables, 2) ability to understand the relationships between data by manipulating data, such as sorting or filtering, 3) knowledge of the basic visualization categories, such as bar chart, line chart, heat map, and geographic map, 4) ability to interpret the graph's scale, legend, and axis, 5) ability to identify the data format (geographic data or numerical data) used in the visualization, 6) ability to extract or interpret meaningful information from visualizations using pattern recognition and mathematical thinking.

In total, nine OGD literacy capabilities were formulated in the stage of *sensemaking*, shown in table 10.

Table 10: OGD Literacy Capabilities in the Stage of *Data Sensemaking*

Behavioral process	Actions	OGD literacy capabilities	Codes
Understand	Understand Context	Ability to understand when, where, and how the identified datasets were collected, who collected them, and what the dataset is about	ASC-20
Understand	Understand Context	Ability to comprehend this context information, and use it to help understand the data	ASC-21
Understand	Understand Fields	Ability to understand the meaning of column headings and/or the corresponding values by using a provided data dictionary or metadata, or based on personal abilities, such as previous background knowledge, and Google searches	ASC-22
Schematize	Understand Table Structure	Ability to build connections between the columns and/or between tables	ASC-23
Schematize	Understand Table Structure	Ability to understand the relationships between data by manipulating data, such as sorting or filtering	ASC-24
Schematize	Understand Visualizations	Knowledge of the basic visualization categories, such as bar chart, line chart, heat map, and geographic map	ASC-25
Schematize	Understand Visualizations	Ability to interpret the graph’s scale, legend, and axis	ASC-26
Schematize	Understand Visualizations	Ability to identify the data format (geographic data or numerical data) used in the visualization	ASC-27
Schematize	Understand Visualizations	Ability to extract or interpret meaningful information from visualizations using pattern recognition and mathematical thinking	ASC-28

5.3.4 The Fundamental Action-specific OGD Literacy Capabilities in the *Data Use* Stage

This section elaborates on the needed fundamental OGD literacy capabilities in the stage of *data use*. As presented in figure 35, we visualized the identified capabilities in the *data use* stage as a whole to compare with the capabilities in the behavioral process level under

other stages. This visualization is mainly because we classified the behaviors under the data use stage into a nuanced level, which caused only a few related posts to be identified. In addition, as stated above, because users may not think an online forum is an applicable place to ask complex skill questions, there are fewer data-use-related questions on the forum posts. In this section, as the process of *preparing data* consists of several granular behaviors, we discussed it first and then explicated other behaviors and their corresponding capabilities together.

5.3.4.1 The Capabilities in the Process of *Preparing Data*. The forum data analysis disclosed 41 posts that can infer the OGD literacy capabilities in the process of *preparing data*. We initially included the ability to extract, import, and clean data when designing the coding schema (section 4.7.2). We then extended the capacities to merge and transfer data based on the forum posts. Ultimately, three codes were utilized to classify the forum posts, including Use-EIC, Use-Trans, and Use-Merge.

Based on the coding schema (section 4.7.2), the code Use-EIC refers to a user being able to extract, import, and clean data using various methods. We identified 25 posts that can manifest the capabilities related to extracting, importing, and cleaning data. E.g., *“Once filters are applied, the table at the bottom of the visualization returns only the relevant records, and you can export that filtered dataset by clicking on the icon of three horizontal lines that are right about the search bar near the table.”* This post demonstrates that this user knew how to filter and extract the needed data from a large dataset. E.g., *“When I tried changing all of the names of the shapefile components, for example, from ‘13bdf129-5a8a-48ae-9582-af09cd1ebd3e2020329-1-1cae11w.6h0he.shp’ to ‘bike-network.shp’ and I was able to use it normally in ArcGIS Pro.”* This post indicates that this user was able to import data (shapefile) to a map tool (ArcGIS Pro.) E.g., *“We worked with this data over the summer and composed a blog post which walks through a process to clean the data. You are correct the data are lacking IDs and have many duplicates.”* This post denotes that this user was able to clean the data for future use.

The code Use-Trans refers to a user being able to convert the data types and convert data files between different data formats. Eight relevant posts were discovered from the

forum post analysis. For instance, *“Rather than re-geocoding all of the addresses, there are 2 columns (‘lat’ and ‘long’) that you can use to create an X/Y event layer and then export as a shapefile.”* This post shows that this user transferred the tabular data to a shapefile.

The code Use-Merge refers to a user being able to merge data from different datasets. We found eight posts that discussed the ability to merge data. The following posts indicate that the users knew how to merge data. E.g., *“You can do a join between the DOR parcel data and the OPA data using the ‘registry_number’ in the OPA data”*. This post displays that this user was aware of the basic concept and rules of merging data, requiring an identical field to merge data. E.g., *“I am analyzing parcel-level data and joining such data to OPA’s universe of parcels.”* This post shows that this user was able to perform a merging data operation.

Interview data analysis detected similar needed capabilities for preparing data. For example, P1 raised that *“being able to clean data is really really important.”* P5 expressed the same idea; specifically, *“we had to know how to connect data and clean it.”* These two examples signify users’ ability to clean data. Participants also proposed the capacity of merging data. For example, P16 stated that *“being able to merge data to create a new table for further data analysis is a crucial step.* In our interview studies, no participants discussed transferring data into different formats or converting different data types. However, several participants argued the essential knowledge of data formats and types. For example, P3 depicted that *“You would also need some types of knowledge of file formats, CSV or XML. A data literate person should have an understanding of a file format.”* P5 put forward the importance of understanding data types, *“I think there is numerical data or text data, so just being able to understand what that means, and then apply it.* These basic concepts are fundamental to transferring and using data.

5.3.4.2 The Capabilities in the Process of *Analyzing data, Representing Results, Interpreting, and Communicating Data.* Forum post analysis identified the capabilities of analyzing data (5), representing results (8), and interpreting data/results (2). The following posts provide evidence for those three needed abilities. E.g., *“For instance, the 10/29 numbers are 4111 Negative and 447 Positive, which by my math would be 4558 Total with a Positivity % of 9.81% (double-checked by percentagecalculator.net) but the ”Positivity*

%” purple line graph for 10/29 lists 7.7%.” This post indicates that this user knew how to analyze data. E.g., *“I want to make a map with parcel-level zoning data (available from OPA) that are rendered in polygons (from DOR).”* This post shows that this user aimed to create a map visualization to plot data and represent the result. E.g., *“Looking at adjacent zones for 2-3 days, you can start to see patterns, and it becomes fairly clear which day the garbage truck is coming to your block.”* This post displays that this user was able to interpret data by observing patterns.

Interview data analysis provides more insights into the OGD literacy capabilities of using data. Several participants proposed the capacity to analyze data. For example, P1 described that *“any statistics that we found, like the percentage of the food garden in a certain area of Pittsburgh was really important, and any kind of numbers we are then comparing with other datasets, our findings from those datasets, so the statistical analysis and comparison is kind of quite an essential part of evaluating your data and working through it.”* P8 expressed that *“I wish I had more math skills and a basic understanding of statistics, like correlation. It would be really useful to know what I am doing here.”* P13 also mentioned that *“OGD literacy requires some kind of college-level statistics and analysis, and with this skill, I don’t think it is too difficult to see generally what is going on with the data.”* These three examples stressed that statistical analysis is important for analyzing data.

Participants also discussed the ability to represent results. For example, P7 stated that *“finding what is the best way to visualize the data, just like giving someone the whole dataset, and showing them what the content is through that visualization is an important literacy.”* P8 also proposed that *“I think visualizing patterns in the data can reveal a lot more about it than just looking at the numbers.”* These two examples argue that being able to visualize results is a critical capability.

Regarding interpreting data or results, participants put forward their ideas. P7 shed light on the importance of *“being able to interpret the statistics you are getting from the data, and not just interpreting the statistics, but interpreting what all of those numbers are actually doing to support for your project or your research.”* P10 considered that *“it is also important for data literacy to be able to understand what the data is saying after you manipulated data; being able to come up with some of the conjecture as to why the data is doing what is doing.”*

These two examples indicate the significance of deriving meaning from data analysis.

Other capabilities in the data use stage were also brought up by participants. For example, knowing how to use Excel (P3, P13, P17) and programming languages, e.g., Python (P10), to manipulate data. P14 proposed that *“dealing with GIS data is very different from tabular data. For example, doing coordinates reference projections, you have to make sure to reproject them to all the same coordinates references. Also, you have to do geometry types, like a polygon.”* According to the analyses of forum posts and interview data, we found many users utilizing OGD to create geospatial maps. Therefore, the knowledge associated with geospatial data and the skills in using related tools need to be considered critical OGD literacy capabilities.

Based on the data analysis, we created the needed fundamental capabilities for the stage of *data use*, shown in Table 11. Because the forum data and interview data did not obviously reveal the capability needed for data communication, we relied on the data literacy-related literature and focus group suggestions. We acknowledge that being able to tell an effective data story is a critical ability for a data task.

Table 11: OGD Literacy Capabilities in the Stage of *Data Use*

Behavioral process	Actions	OGD literacy capabilities	Codes
Prepare	Extract Data	Ability to extract the needed granular data (e.g., selecting a subset of rows and columns from a data table)	ASC-29
Prepare	Clean Data	Ability to clean the data through different approaches (identifying anomalies, removing duplicate records, fixing errors, and normalizing values), manually, through a tool (like OpenRefine), or through programming	ASC-30
Prepare	Import Data	Ability to import/upload data to a GIS tool or a programming platform	ASC-31
Prepare	Convert(types or formats)	Knowledge of basic data types: integer, float, string, boolean, character	ASC-32
Prepare	Convert(types or formats)	Knowledge of basic data structure for different data formats (e.g., CSV and JSON)	ASC-33
Prepare	Convert(types or formats)	Ability to convert the data types	ASC-34
Prepare	Convert(types or formats)	Ability to convert data between different data formats	ASC-35
Prepare	Reshape Data	Ability to aggregate tables of data (e.g., summing columns or creating pivot tables)	ASC-36
Prepare	Reshape Data	Ability to transform data tables to achieve the needed form (e.g., by converting between narrow and wide tables)	ASC-37
Prepare	Merge Data	Knowledge of the basic concept and rules of merging data, such as the need for finding a common field to match on	ASC-38
Prepare	Merge Data	Ability to merge data through different methods, such as tools (Excel) or programming languages	ASC-39
Analyze	Analyze Data	Ability to select the proper analysis method to analyze data in terms of the proposed questions, such as descriptive statistics, exploratory data analysis, and confirmatory data analysis	ASC-40
Analyze	Analyze Data	Ability to analyze data by using statistical/mathematical knowledge and methods (percentage and correlation)	ASC-41
Analyze	Sort Data	Ability to sort data according to the needs	ASC-42
Analyze	Filter Data	Ability to filter data according to the needs	ASC-43
Analyze	Analyze Data	Ability to avoid data analysis errors, e.g., misbeliefs (confusing correlation and causation)	ASC-44
Represent results	Plot Data	Ability to choose appropriate forms of visualizations for the data	ASC-45
Represent results	Plot Data	Ability to create basic visualizations (tables, graphs, and charts) to present data	ASC-46
Represent results	Plot Data	Ability to create geospatial maps to visualize data	ASC-47
Interpret	Distill Meaning	Ability to interpret the patterns or the analysis results	ASC-48
Interpret	Distill Meaning	Ability to draw meaning from results by integrating related contextual information to help make decisions	ASC-49
Communicate	Tell Data Stories	Ability to deliver an informative and easy-to-understand data story, pitch at the correct level for the target audience	ASC-50

5.3.5 The Fundamental Action-specific OGD Literacy Capabilities in the *Data Sharing* Stage

The behavior of *sharing* data was mainly revealed from the forum data analysis (9, 1.7%). For instance, “*Would be glad to share my copy with anyone if there is some convenient way to transmit the file. The Excel file contains 58,000 rows, which is a little over 300 MB.*” This post embodies that the user was willing to share the collected data with others, provided a simple description for the data size and format, and then asked for an appropriate sharing platform. Another similar post shows that “*I keep a history of the daily COVID-19 data published by the City. You can find that here: <https://github.com/ambientpointcorp/covid19-philadelphia>.*” This post reflects that this user was also willing to share the collected data with others, and directly provided the downloading link for the data, which indicates that this user knew how to use platforms to share data.

Interview participants expressed that they shared the task results with others but did not share the originally collected data. For example, they would share a report for the project or the maps they created with others. In the first focus group, we brought up the behavior of *sharing data*, and one expert suggested that creating metadata or a data document for the data is necessary to facilitate other people’s understanding of it. Synthesizing all the results, we develop three capabilities for sharing data, shown in Table 17.

Table 12: OGD Literacy Capabilities in the Stage of *Data Sharing*

Behavioral process	Actions	OGD literacy capabilities	Codes
Share	Document Data	Ability to describe the processes used to generate the output so that the work can be understood and reproduced (e.g., the descriptions of methodologies for collecting and processing data, commenting code, a laboratory notebook, and limitations of the data)	ASC-51
Share	Document Data	Ability to create data descriptions for data to be shared (e.g., metadata or data dictionary)	ASC-52
Share	Share Data	Ability to upload and edit data on different platforms (e.g., Google Sheets and GitHub)	ASC-53

5.3.6 The Fundamental Action-holistic OGD Literacy Competencies

The Action-holistic OGD Literacy Competencies are defined as the competencies needed prior to carrying out a task, such as a certain level of civic data domain knowledge and basic concept of data, or the competencies involved in multiple behavioral processes, such as critically examining data and locating help. These findings were not included in the initial coding schema; they were partially observed in the analyses of forum posts and interview data but primarily based on the focus groups. The two focus groups encompass the faculty members who teach data science classes and the OGD portal’s staff who organize OGD literacy workshops or directly answer users’ questions. Instructors can identify the questions students or users have and know the needed knowledge or skills to solve user issues. Therefore, their observations of the required competencies for OGD literacy are from higher-level perspectives and enrich the taxonomy of the OGD literacy capabilities.

Three categories of OGD literacy competencies beyond the individual behavior in the H-OGD-I model were identified, including conceptual foundations, advanced actions, and dispositions. This section illustrates the three categories.

5.3.6.1 Conceptual Foundations. In this dissertation, conceptual foundation refers to the fundamental concepts required for performing a data task, including understanding the basic concepts of data and civic data domain knowledge. For example, one expert put forward that *“what is data? The concept of what a data atomic unit is and how data is divided; if they don’t have those skills, then they are not going to be able to engage in this activity.”* Another expert stated that *“there is some domain understanding is requisite in order to understand what the data means. Highlight the domain understanding and domain knowledge is necessary.”* Interview participants also raised that knowing basic data concepts (P2, P6, P16), e.g., data types and data formats, and basic civic data domain knowledge (P3, P14), e.g., the existing data sources, are requisite for working on a data task. The specific competencies are shown in table 13

Table 13: Competencies-Conceptual Foundations

Categories	Actions	OGD Competencies	Codes
Conceptual Foundations	Understand Basic Concept	Ability to understand the basic concepts of data. For example, what is data?	AHC-1
Conceptual Foundations	Understand Basic Concept	Awareness of how essential domain knowledge is to accomplishing a data task (e.g., domain knowledge “reduces the cognitive load” of a data task)	AHC-1

5.3.6.2 Advanced Actions. The competencies are named advanced actions because we considered these abilities advanced competencies that would enable users to efficiently execute their data tasks and then successfully complete them.

Being reflective indicates that users can keep reflecting on what they already know and learn in practice and apply the newly obtained knowledge to the project. For example, one expert argued that *“you should be able to self-reflect on what you have done, criticize your own thinking.”* We also considered that being reflective is being able to keep thinking back through one’s own work and processes. For instance, find inconsistencies or opportunities for improvement. The specific explanations are presented in table 14.

The competency locating help denotes that users have the ability to figure out what help is needed at different process points and find contact information through available metadata information or have related background information to locate the contact person for particular datasets. The forum data presents that users locate help through metadata (Contact-Metadata = 6), e.g., *“these are great questions. I don’t know the answer, but you might also contact the email address that is listed with the City’s COVID datasets: CO...@phila.gov,”* or through personal background information (Contact-RBI = 11). E.g., *“The Department of Public Health provides this: If your question is not answered on the web page, contact the office of the epidemiologist.”*

We also determined that the competency of locating help enables users to acquire civic domain knowledge through various channels (e.g., workshops, forums, news, and asking

related practitioners). It also includes understanding about broader background information for datasets, including the origin of a dataset, dataset stories, or information about data publishers. For example, P12 found the needed data because s/he attended a WPRDC workshop and then had the chance to seek help. A forum post shows that “I’m not aware of anyone trying to get access to the XXX data, but I frankly don’t know. You could potentially email the OpenDataPhilly mailing list.” This post suggests the data seeker can locate help from an email list.

Applying domain knowledge signifies that users are able to integrate domain knowledge to perform a data task (table 14). This competency is a critical point during the entire process. Forum data analysis identified that many questions required to apply broad domain knowledge to be answered (Context-Overall = 28). E.g., *Bench data would be outdated quickly since benches often are damaged and removed. Memorial plaques attached to them often also get removed.* Interview participants also propose the same idea. For example, P13 described that *“you need to have some sense of understanding, like public health. And if you don’t, it might give a very skewed perception of what is actually happening. And there is nothing about that in the data, obviously. Just the raw data itself doesn’t really tell you the full story, so understanding that is very important.”*

Table 14: Competencies-Advanced Actions

Categories	Actions	OGD Competencies	Codes
Advanced Actions	Reflect	Ability to reflect on newly obtained feedback/results and apply them to the project practice	AHC-3
Advanced Actions	Reflect	Ability to think back through one’s own work and thought processes, find inconsistencies or opportunities for improvement, and optionally backtrack to repeat an earlier stage in the taxonomy/workflow, making different choices the second time	AHC-4
Advanced Actions	Locate Help	Ability to figure out what help is needed in different processes	AHC-5
Advanced Actions	Locate Help	Ability to find contact information through metadata information or have related background information to locate the contact person for particular datasets	AHC-6
Advanced Actions	Locate Help	Ability to acquire civic domain knowledge through various channels (e.g., workshops, forums, news, and asking related practitioners) and broader background information for datasets, including the origin of a dataset, operational information about how the data was collected, some stories for the dataset, or some information about data publishers	AHC-7
Advanced Actions	Apply Domain Knowledge	Ability to integrate domain knowledge to perform a data task, such as framing data-centric questions, deeply understanding data, and telling a comprehensive data story	AHC-8

5.3.6.3 Dispositions. In this dissertation, *disposition* refers to users’ ability to form mindsets that will enhance equity and equality in data-driven tasks, including critically examining data, ethically using data and operating tasks, as well as avoiding or reducing biases.

Interview participants and focus group experts raised the idea of how to examine data critically. For example, P14 claimed that “*always remember that the people who make the data in the first place are not perfect. They are just regular people. As a result, there are to be some mistakes in the data in some ways.*” One expert emphasized that some biases may have shaped that data. This expert explained the importance of “*critically considering the*

creator and how that shapes the way that you should be understanding and thinking about the data. I mean, the creator itself is really important for somebody to consider when they are looking at a dataset: there is a critical lens assigned to thinking about the data producer, and that impacts the way we should understand and interpret this data.”

The ethical use of data has been discussed long before the open data movement started. One expert supplemented the importance of the ethical use of data by stating that *“not only following the laws of political dangers but also, being aware of and acknowledging the biases within data because almost every dataset does have that. I think that will be part of like ethical use is making sure that the users or the consumers are aware of what biases may be inside of those data sets.”* We incorporated this suggestion into the capabilities for sharing data, which not only includes sharing the data itself but also contains sharing the data collection method and any author-identified data limitations.

Experts also proposed the idea of avoiding non-conscious biases or misunderstandings about data, such as confirmation bias or some misunderstanding of the statistics. For example, *“correlation is not causation.”*

Based on these analyses, we concluded six essential competencies within three actions for dispositions, as presented in table 18

Table 15: Competencies-Dispositions

Categories	Actions	OGD Competencies	Codes
Dispositions	Critically Examine Data	Ability to understand the limitations of data. For example, knowing that data is not perfect and data is not absolutely objective	AHC-9
Dispositions	Critically Examine Data	Ability to locate hidden gaps/biases/limitations	AHC-10
Dispositions	Critically Examine Data	Ability to apply a critical lens to examine the biases in the data collection process that may have shaped that data, e.g., the trustability of sources, systemic bias, or selection bias	AHC-11
Dispositions	Operate Ethically	Ability to use data ethically, such as by following related laws, policies, and standards (e.g., the dataset’s license), as well as avoiding disclosing people’s private information	AHC-12
Dispositions	Avoid or Reduce Biases	Ability to reduce non-conscious biases, e.g., confirmation bias and selection bias	AHC-13
Dispositions	Avoid or Reduce Biases	Ability to avoid misusing or misrepresenting data, e.g., cherry-picking data that supports one’s desired result), that is, demonstrating scientific integrity	AHC-14

5.3.7 The Evaluation of the Taxonomy for OGD Literacy Capabilities

We initially identified the needed fundamental OGD literacy capabilities through the two empirical studies. To better validate and refine these findings, we then conducted two focus groups with experts in the fields of OGD literacy and data science. Four experts were invited to attend our first focus group. Three of them are faculty members in a university in the U.S.; two of them teach data science courses by using OGD, and another faculty teaches a course introducing OGD and metadata. The fourth expert works for a local OGD portal, with rich experience working with OGD users. The second focus group embraced three experts. Two of them are faculty members in two different universities in the U.S. They both have experience in researching and teaching OGD literacy. The third expert has been working for a local OGD portal for several years and has conducted many workshops

associated with OGD literacy.

Three primary criteria were applied to guide this evaluation: Correct, Clear, and Comprehensive. Specifically, *correct* refers to whether or not the explanation of the capabilities is correct and makes sense to the experts. *Clear* indicates whether or not the explanation of the capabilities is clear to the experts. *Comprehensive* denotes whether or not the components of the capabilities are comprehensive. The taxonomy was refined after the first focus group, which means the taxonomy that the second focus group reviewed is an updated version based on the suggestions from the first focus group. Fewer suggestions were brought up in the second group. Therefore, we considered this taxonomy was assessed sufficiently.

The experts in the first focus group proposed many crucial suggestions for this taxonomy. They brought up that the nature of a data task is interactive and repeatable, which is consistent with our H-OGD-I model. Therefore, being reflective, critically examining data, and avoiding biases are essential. In fact, most of the competencies within *advanced actions* and *depositions* were inspired and created by their suggestions. The specific application of their ideas is illustrated in the result section. The second focus group assessed the refined version of the taxonomy. As a result, three experts agreed that the components of this taxonomy are correct, clear, and comprehensive, and they think this taxonomy will be valuable in practice and research studies.

5.3.8 Mapping to Existing Data Literacy Capabilities

In this dissertation, we mapped this new taxonomy for OGD literacy capabilities to the data literacy skills that were developed by existing scholarly studies, shown in table 16. We chose three research studies to summarize the existing data literacy capabilities based on two primary reasons. First, all three research studies conducted a thorough literature review of the scholarly works regarding information and data literacy and then proposed a clear set of data literacy capabilities. Second, the three research works reviewed data literacy from different angles. As discussed in the literature review section, information literacy was developed long before data literacy, and data literacy is considered an extension of information literacy. Prado & Marzal (2013) proposed the data literacy competencies according to in-

specting information literacy capabilities and incorporating them with data literacy (shown [1]). Wolff, Gooch, et al. (2016) discussed the data literacy capabilities for four groups: research focus, classroom focus, carpentry focus, and inclusion focus (general public use) (shown [2]). Matthews (2016) proposed data literacy skills to address the needs of citizens in today's society (shown [3]). These three research studies cover the data literacy capabilities that originated from information literacy and the abilities associated with citizens' needs, which are closely connected to our OGD data literacy capabilities.

Table 16: OGD Literacy Capabilities Mapping

NO.	Capabilities	Related works	Mapping Capabilities
1	Ability to know what is meant by data	[1]	AHC-1
2	Ability to to frame a research question or problem that can be addressed with application of data	[3]	ASC-4,ASC-5,ASC-6
3	Ability to understand the role and impact of data in society in different contexts	[1] [3]	N/A
4	Ability to recognize source data value, types and formats	[1]	ASC-32
5	Ability to determine when data are needed	[1] [2]	ASC-7
6	Ability to access data sources appropriate to the data needed	[1] [2] [3]	ASC-8,ASC-9
7	Ability to critically assess data and their sources (quality and credibility)	[1] [3]	ASC-16
8	Ability to detect when a given problem or need cannot be (totally or partially) solved with the existing data and, as appropriate, undertake research to obtain new data.	[1]	ASC-7
9	Ability to collect and acquire data	[2] [3]	ASC-18,ASC-19,ASC-29
10	Ability to clean, transform, and merge data	[3]	ASC-30,ASC-32,ASC-33,ASC-34,ASC-35,ASC-38,ASC-39
11	Ability to handle and analyze data	[1] [2] [3]	ASC-40,ASC-41,ASC-42,ASC-43,ASC-44
Continued on next page			

Table 16 – continued from previous page

NO.	Capabilities	Related works	Mapping Capabilities
12	Ability to present quantitative information (specific data, tables, graphs, in reports and similar)	[1] [2] [3]	ASC-45,ASC-46
13	Ability to use data ethically	[1]	AHC-12
14	Ability to apply results to learning, decision making or problem-solving	[1] [2][3]	ASC-49
15	Ability to Interpret information derived from datasets	[2] [3]	ASC-48
16	Ability to preserve and archive data for reuse	[3]	ASC-51,ASC-52
17	Ability to plan, organize and self-assess throughout the process	[1] [2]	AHC-4
18	Ability to communicate outcomes	[2] [3]	ASC-50
19	Ability to contextualize domain knowledge	[2]	AHC-2,AHC-8

[1] Incorporating data literacy into information literacy programs: Core competencies and contents (Prado & Marzal, 2013)

[2] Creating an understanding of data literacy for a data-driven society (Wolff, Gooch, et al., 2016)

[3] Data literacy conceptions, community capabilities (Matthews, 2016)

The three research studies enable us to summarize 19 data literacy capabilities. Of the 53 action-specific OGD literacy capabilities (ASC), only 28 of the 53 were mapped to the existing data literacy capabilities. Also, with a total of 14 action-holistic competencies, only 5 of the 14 were classified into the summarized data literacy capabilities. Our comprehensive OGD literacy capabilities can be mapped to most of the summarized capabilities except one - the ability to understand the role and impact of data in society in different contexts. After thorough consideration, we decided not to include this capability in our proposed OGD literacy capabilities. The main reason for making this final decision is that our data analysis from the three data sources and focus groups did not reflect this particular ability, and we view it as a motivation to perform a data-driven task.

Another two significant discoveries were observed after mapping these capabilities. The first discovery is that our new taxonomy is more detailed than other studies. Because the

capabilities were broke down based on user behaviors, we found that some data literacy capabilities correspond to multiple OGD literacy capabilities proposed in this dissertation. For example, in order to prepare data, existing studies simply described it as the “ability to clean, transform, and merge data.” However, these proposed new OGD literacy capabilities pointed out comprehensive knowledge or skills to perform activities. For example, to be able to transform data, users need to have the knowledge of basic data types: integer, float, string, boolean, character (ASC-32) and basic data structure for different data formats (e.g., CSV and JSON) (ASC-33), and have the ability to convert data types (ASC-34) and formats (ASC-35). Also, “the knowledge of the basic concept and rules of merging data” (ASC-38) and the “ability to merge data through different methods, such as tools or programming languages” (ASC-39) were depicted as the same capability in the summarized list. Therefore, we have seven distinct capabilities which correspond to the one in the summarized list. In addition, some of the OGD capacities not covered by the summarized data literacy capability list also displayed the details. For instance, the data literacy list contains the “ability to access data sources appropriate to the data need.” This capability indicates the ability to find and access data. Our OGD literacy capabilities also proposed the specific skills of seeking data, such as choosing the right keyword (ASC-10) and knowing data categories (ASC-12).

Using the H-OGD-I model to observe the corresponding needed OGD literacy capabilities is a user-centered method that can detect a comprehensive set of abilities that enable users to efficiently use data. The second essential discovery is that our OGD literacy taxonomy revealed additional critical capabilities through observing users’ OGD behaviors, including action-specific OGD literacy capabilities and action-holistic competencies. Regarding action-specific OGD literacy capabilities, the capabilities of making sense of data (ASC 20-24) were not listed in previous works, but they are critical in the OGD literacy capabilities, e.g., the ability to build connections between the columns and/or between tables. Furthermore, we identified the new behavior of sharing data; therefore, the corresponding capabilities were developed, e.g., the ability to upload and edit data on different platforms (ASC-53). Additionally, documenting data was classified under the behavior of sharing data, whereas the summarized data literacy capabilities described it as preserving or archiving data. In addi-

tion, most action-holistic competencies (9 of the 14) in this dissertation were not contained by the summarized data literacy list, e.g., “ability to reflect on newly obtained feedback/results and apply them to the project practice” (AHC-3) and the three competencies in critically examining data. Those are essential abilities to perform a data task.

These comprehensive and granular OGD literacy capabilities and competencies provide a clear direction for designing data literacy materials. These capabilities and competencies provide a framework for understanding the core skills and knowledge needed to work with data. They also provide a roadmap for developing materials that can help people understand the fundamentals of data literacy and how to use data to make informed decisions. By having a clear direction for designing data literacy materials, governments and public sectors can ensure that their employees, partners, and users have the knowledge and skills they need to work with data effectively.

5.3.9 Outcome 2: The Taxonomy for OGD Literacy Capabilities

This section demonstrates the taxonomy of OGD literacy capabilities in each H-OGD-I stage. The symbols in figure 36 represent the fundamental Action-holistic OGD Literacy Competencies that could be needed throughout the whole H-OGD-I process.








Categories	Actions	OGD literacy competencies	Symbol
Conceptual Foundations	Understand Basic Concepts, and Aware domain knowledge importance	1. Ability to understand the basic concepts of data. For example, what is data? 2. Awareness of how essential domain knowledge is to accomplishing a data task (e.g., domain knowledge "reduces the cognitive load" of a data task).	
Advanced Actions	Reflect	1. Ability to reflect on newly obtained feedback/results and apply them to the project practice. 2. Ability to think back through one's own work and thought processes, find inconsistencies or opportunities for improvement, and optionally backtrack to repeat an earlier stage in the taxonomy/workflow, making different choices the second time.	
	Locate help	1. Ability to figure out what help is needed in different processes. 2. Ability to find contact information through metadata information or have related background information to locate the contact person for particular datasets. 3. Ability to acquire civic domain knowledge through various channels (e.g., workshops, forums, news, and asking related practitioners) and broader background information for datasets, including the origin of a dataset, operational information about how the data was collected, some stories for the dataset, or some information about data publishers..	
	Apply Domain Knowledge	Ability to integrate domain knowledge to perform a data task, such as framing data-centric questions, deeply understanding data, and telling a comprehensive data story.	
Dispositions	Critically Examine Data	1. Ability to understand the limitations of data. For example, knowing that data is not perfect, and data is not absolutely objective. 2. Ability to locate hidden gaps/biases/limitations. 3. Ability to apply a critical lens to examine the biases in the data collection process that may have shaped that data, e.g., the trustability of sources, systemic bias, or selection bias.	
	Operate Ethically	Ability to use data ethically, such as by following related laws, policies, and standards (e.g., the dataset's license), as well as avoiding disclosing people's private information.	
	Avoid or Reduce Biases	1. Ability to avoid or reduce non-conscious biases, e.g., confirmation bias and selection bias. 2. Ability to avoid misusing or misrepresenting data, e.g., cherry-picking data that supports one's desired result), that is, demonstrating scientific integrity.	

Figure 36: The Fundamental Action-holistic OGD Literacy Competencies

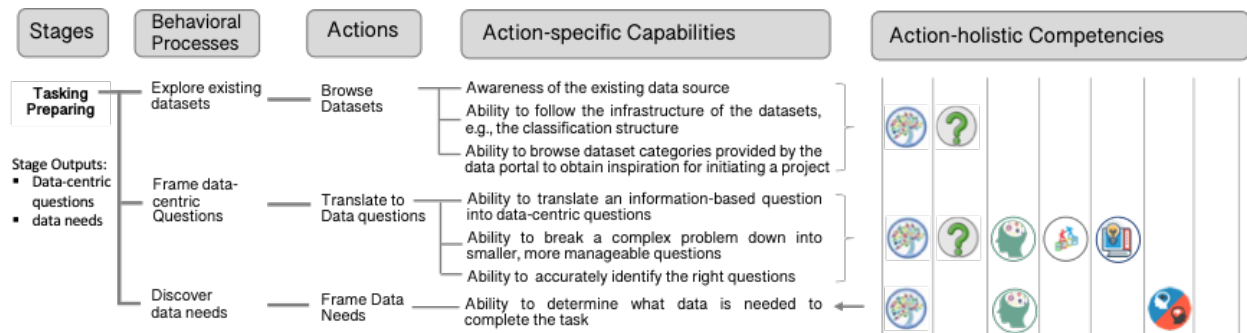


Figure 37: The Taxonomy for OGD Literacy Capabilities - Task Preparation

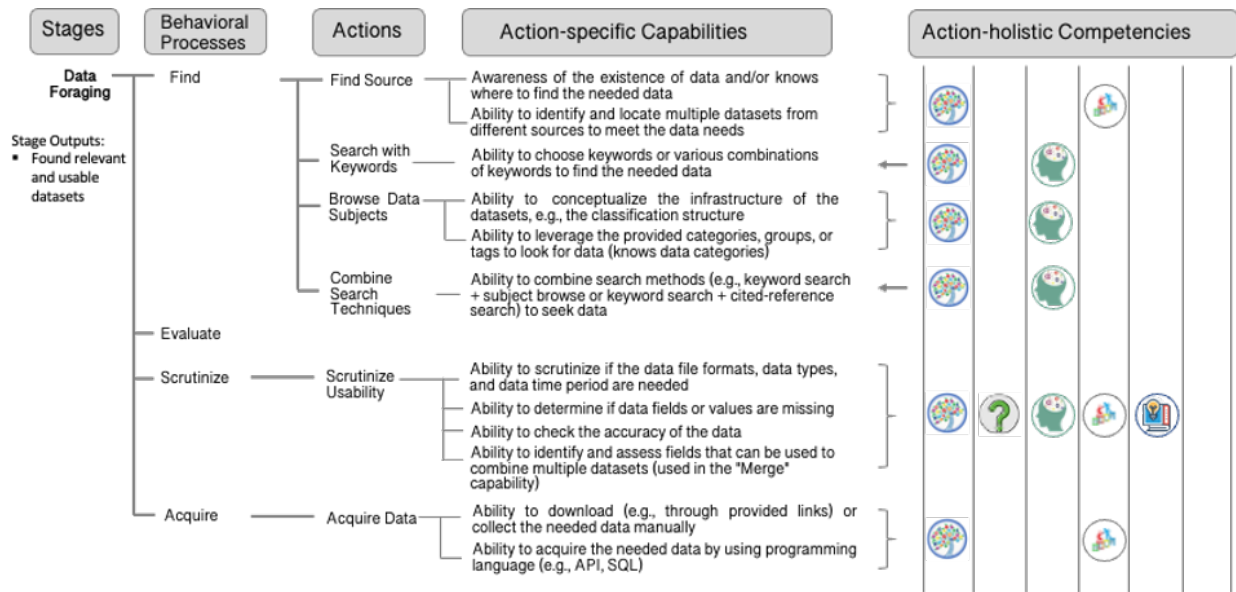


Figure 38: The Taxonomy for OGD Literacy Capabilities-Data Foraging

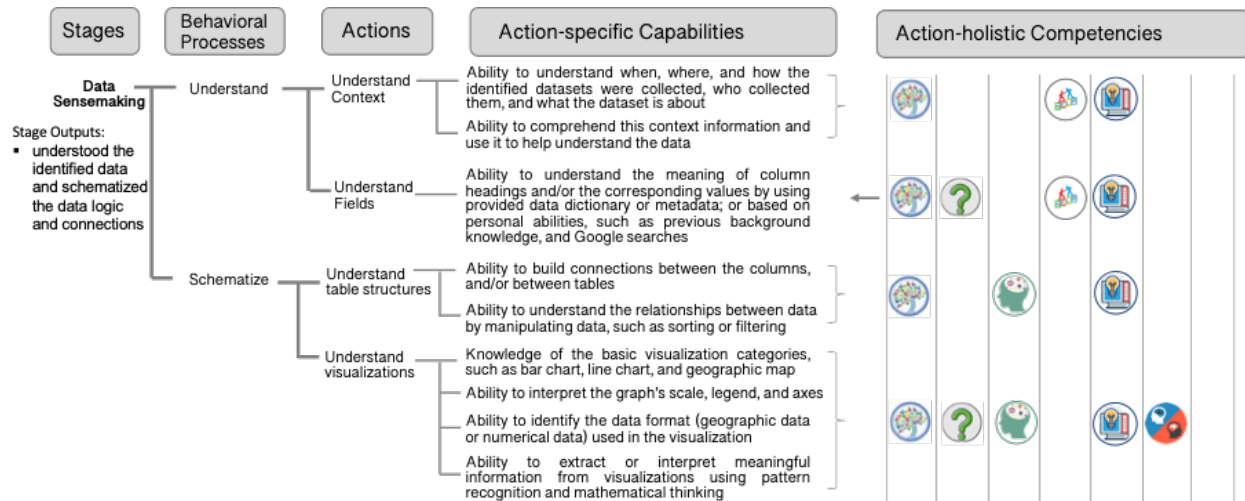


Figure 39: The Taxonomy for OGD Literacy Capabilities-Data Sensemaking

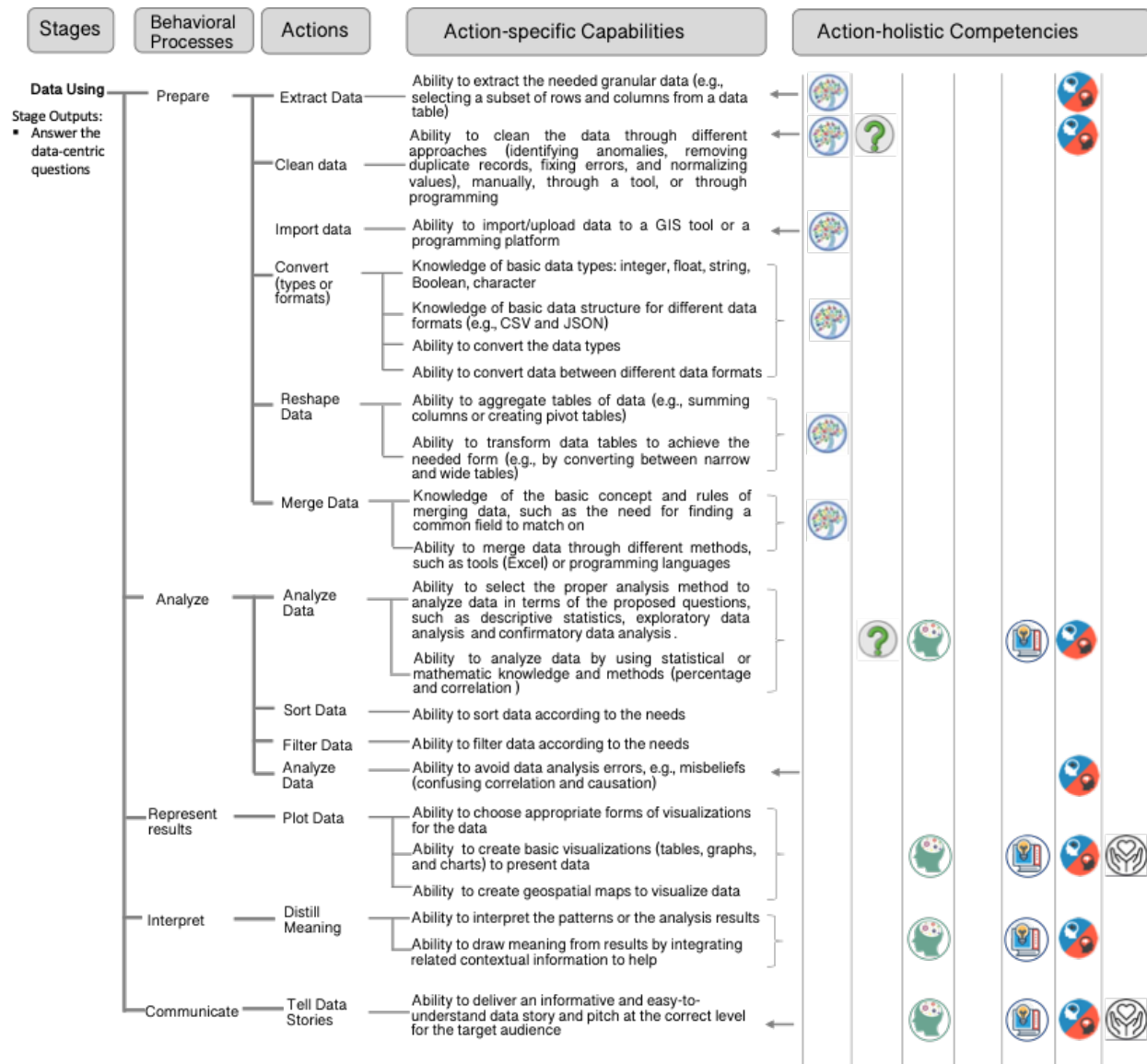


Figure 40: The Taxonomy for OGD Literacy Capabilities - Data Use

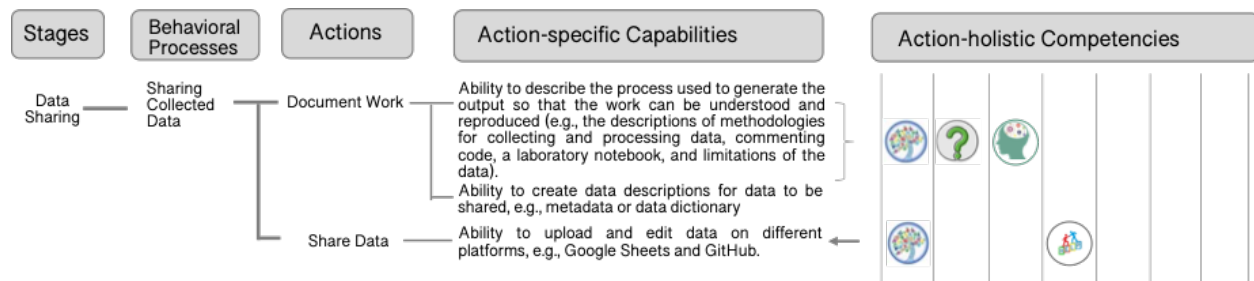


Figure 41: The Taxonomy for OGD Literacy Capabilities - Data Sharing

6.0 Discussion and Implications

6.1 Discussion of Results

6.1.1 OGD Users' Accessing Behavior Patterns

Channels. The results from Study 1 show that although the channel of organic search is not the most commonly used channel for all the portals, with successful access and a conversion rate, organic search still plays an essential role in leading people to the portals. This finding suggests that OGD portals may benefit from finding ways to get a greater amount of exposures on search engines, which could improve the accessibility of OGD. Furthermore, referrals are a good way to advertise datasets. However, better collaboration with the referral sites in order to generate a better success rate for referral accesses is needed. For example, the links from the referral sites should be more specific to certain datasets so that the users are better guided in the referral. Also, the organizers of OGD portals should be able to see all the referral sources, which means they can observe which sources gather more or fewer users, and then investigate these results. For instance, when they see one source has several users but few of them download the dataset, they may want to examine the reasons for this phenomenon. Or, the OGD organizers can examine why a channel draws fewer users, considering the relevance of the topic to the source, or broken links, for example. Through checking in this data, OGD portals could better understand themselves and their users, which can be conducive to the governance of the datasets with a low cost. Moreover, given the fact that users use various methods to find the OGD portals, the organizers of the portals may check their own situations to find out their weakest points and improve them.

The results on channel preferences are different from some existing studies that examined national-level OGD portals. Kacprzak et al. (2019) found that the majority of their users prefer to use search engines to locate the portals. Also, Koesten et al. (2017) presented that 18 out of their 20 participants often used Google to find data online. However, a recent study (Ibáñez & Simperl, 2022) argued a similar result that local-level OGD portals

present less access through organic search and much more through direct access. They consider that is because the national-level OGD portals hosted a large number of datasets, which leads to greater chances of exposure to a web search engine. In addition, considering that the three OGD portals we studied have different most used channels, maybe it is not surprising to see the dissimilarity among users' behaviors in different portals. This phenomenon further tells us that each portal may have its own characteristics, including its users, operation management approaches, and system design, so the study of users' access might be contextualized to a specific portal.

The preference of browsing and keyword searching. All three portals' users are found to prefer browsing over searching in accessing the datasets, and we could not explain the patterns only based on the transaction log data. When conducting Study 3, by seeking answers from interview participants, we found three primary possible reasons for this pattern. First, we identified the data-inspired task type, which starts from browsing datasets. Second, interview participants described that OGD portals had done a great job in supporting website navigation and/or topic classification of topics. Therefore, it is easy for them to seek data by simply browsing the categories of datasets. Third, one participant mentioned that the search function in the OGD portals is not as good as the Google search engine. For example, it may be difficult to form a query to search a dataset due to the limited description of the dataset. Or, it is hard to retrieve the data when the needed data is part of one dataset, and the issued queries are not included in the dataset title or description. However, *browsing* could help users narrow down their needs through categories and multi-filter functions. All these challenges or preferences could lead to more browsing behaviors than keyword searching behaviors.

On the other hand, some interview participants prefer to use a keyword searching approach to find data, especially when they know what search queries should be used. Therefore, we still think keyword searching is an efficient way to locate data, then the search function in the portal needs improvement. The search query recommendations could be a good way to improve accessibility. The importance of the preferred behavior in OGD users' access motivates us to further examine the mismatch between the topic classification in the portal and users' various needs.

6.1.2 OGD Users’ Challenges When Interacting with Data

The Challenges of Finding Data. Study 4 revealed the most asked question on the online forum concerns *Finding* data. This finding is consistent with prior studies (Lammerhirt, 2017; Swamiraj & Freund, 2015) that point to the poor findability of OGD as a big challenge OGD adoption. Study 4 also pointed out four primary granular data needs that users tried to seek, including data source, data time, data format, and data granularity. This observation matches the findings from the work of Kacprzak et al. (2019) that many search queries contain the detailed descriptions of *geospatial, temporal, and file or data types*. As the detailed *finding* challenges are disclosed, we encourage future studies that 1) investigate the most used data formats OGD users expect to use and the most popular tools OGD users use for processing data, and 2) explore how to store and describe geospatial data to satisfy users’ data needs.

The Challenges of Evaluating Data. In Study 4, we found that no posts asked about evaluating if a dataset is topic-relevant to the data needs. Interview participants also did not propose the challenge of evaluating data. Therefore, we concluded that as long as the title and description of a dataset are clear, no challenges in this behavioral process and, thereof, no corresponding capabilities were proposed. However, a recent study that employed an observational user study presented the challenge of “selecting the most appropriate dataset” (Li et al., 2022). This study described that 9 of 14 participants (64%) selected the dataset where the description clearly stated the needed information was not contained. Five participants (36%) chose other irrelevant datasets. There is no explanation regarding why the participants made that choice. We understand that sometimes people’s thoughts are inconsistent with their actions, which may cause an inaccurate response. For example, when our interview participants were answering questions, they may have forgotten details and thought no barriers were encountered when evaluating the data. In future studies, I will attempt to use the observational user study method to observe users’ evaluating behaviors and explore the pertinent capabilities.

The Challenges of Using Data. Existing studies discovered challenges when users use OGD, such as data quality issues and lack of use standards (Ruijter & Meijer, 2020). The

findings in this dissertation confirmed the previous results and supplemented more details. A significant observation of the *data use* challenge detected many users adopting OGD for tasks of creating geospatial maps or visualizations. Therefore, the challenges they encountered were more about using relevant geographic software, such as importing data to ArcGIS and extracting data from Carto. This usage could be a characteristic of OGD, using geographic data to tell a story. Given this characteristic, three further studies need to be done. The first research direction from the data management perspective is to develop standards for describing geospatial data. If the dataset includes geospatial data or a geographic map, a standardized description for the data that can be applied across all local-level OGD portals would greatly assist users in understanding and using data. The second research direction from the data management perspective is that researchers or OGD administrators may study the effective way of hierarchically managing and preserving data. For example, participants mentioned they attempted to compare the data at a city level, but the existing data were only at a neighborhood level, so they had to alter it to a neighborhood level to conduct the task using zip codes. The third study direction is from the geographic software development perspective. The software designer needs to create a practical approach to enable users to import OGD into the software and make it easy to be used. For example, create a standard data format description, which can be developed with the OGD administrator, to make the process smooth.

6.1.3 The Fundamental Capabilities for OGD Literacy

Study 5 discovered the OGD literacy at the Action-specific level, including 53 capabilities, and at the Action-holistic level, consisting of 14 competencies; in total, 67 capabilities were identified. These capabilities were explored from the user-centered perspective by examining them according to user OGD behaviors. The proposed capabilities show us that OGD literacy requires knowledge and skills from various disciplines, such as domain knowledge, statistical skills, and computational skills. The domain knowledge involves not only civic data domain knowledge but also covers civic knowledge or discipline domain knowledge, such as public health and environment. One expert in the focus group came up with the idea that “*the*

need for civic knowledge should be expanded to knowing city operations, county operations and how government works and stuff like that.” Another expert followed up on this proposal and expressed that *“some domain understanding is requisite in order to understand what the data means. Highlight the domain understanding and domain knowledge is necessary.”* As presented in the result section, one interview participant explicitly claimed that *“I think it is impossible to know this unless you have a background in public health”(P13)*. This finding is consistent with Gray et al. (2018)’s perspective that *“...the ability to account for, intervene around and participate in the wider socio-technical infrastructures through which data is created, stored and analyzed.”* This is a high requirement for non-expert users.

With the rapid development of technologies and the large data size, statistical skills or math knowledge alone are insufficient; using tools or programming languages to process data is a trend for being a data literate person. Consequently, the related capabilities were revealed. Furthermore, in addition to the specific knowledge and skills, holistic competencies are critical for a data task, namely ethically using data, critically examining data, and being reflective in the entire process. The related training or courses can only teach the basic concepts; to master these competencies need to be trained through numerous practical experiences. Based on the analysis of these capabilities, five dimensions of knowledge and skills were proposed for further assessing users’ OGD literacy in the future studies, constructing OGD domain knowledge, information literacy, computational thinking, computational skills (statistic, tool, programming languages), and data communication. As discussed in the section of *The Challenges of Using Data*, geospatial data literacy is starting to be considered as an element of OGD literacy, which would involve many additional and very different skills.

All the observations and analyses indicate that data literacy is ever-evolving in response to the dynamic nature of technology and social environment. Consequently, reducing the gap between non-expert OGD users and expert users may be a challenging task, but it is nevertheless achievable. It is important to note that data literacy is an ongoing process that requires continuous learning and adaptation to the changing environment. As technologies and social environments continue to change, it is important to ensure that there are resources and tools available to help bridge the divide between novice and expert users.

6.2 Discussion of Research Design

The intent in using the mixed method design is to bring together the different strengths of both a quantitative approach (e.g. large sample size, patterns, trends, and generalization) and a qualitative approach (e.g. flexible and in depth). Using this design which will enable me to have a fuller understanding of the research questions as well as to provide various ways to clarify the results from this study. The following sections detail the three research approaches and discuss the rationales of adopting them for the dissertation.

A transaction log is a large-scale data that records user behaviors of interacting with systems, websites, or platforms (Jansen, 2006). Its defining feature is that it captures actual behaviors in real-time, rather than recalled behaviors (Dumais et al., 2014). Transaction log analysis enables researchers to understand the patterns of interacting behaviors. For instance, how people visit/revisit web pages over time and how people search and navigate on a website. Studying transaction logs allow us to unobtrusively examine how users interact with the local OGD portals' online platforms in real situations. One of the advantages of transaction log analysis is that it can reveal the common behavior patterns through large groups of users. However, the disadvantage of this analysis method is that the insight of users, for example, why a user has a certain behavior, cannot be obtained from transaction logs. Therefore, the research methods of content analysis and interviews will be applied to fill the gaps.

As transaction log analysis, online forum posts (content) analysis is also an unobtrusive research method (Babbie, 2001, p.304). Adopting online forum data to study users has been broadly used in the field of the health community (Cui & Lee, 2022; Yang et al., 2019). This is because online forum posts constitute abundant and unobtrusive interactions among users, where they can post requests, answer questions, comment on others' posts, or simply socialize with other users; at the same time, the forums also provide rich user-generated content that can reveal users' motivations, challenges, emotions, stories or interacting behaviors. This unobtrusive characteristic of the online forum data is particularly desirable for studying users characteristics and behaviors because it successfully avoids the worries of the unnaturalness of participants' behaviors in controlled experimental settings. In addition, by using the

forum posts (content) analysis method, researchers can obtain a large scale of data that has descriptions for a topic; the descriptions could be insights from people's experiences of interacting with objects. The forum posts (content) analysis method offers researchers the advantages of using large scale data that encompasses user insights to understand users. However, the exploration will be limited to the existing content; we cannot reveal users' hidden behaviors and considerations behind the behaviors. Therefore, to seek answers for further questions, the interview method is adopted.

The compelling strength of the interview method is its ability to "go deep." Researchers can explore a wide range of insights about an issue or a topic by asking questions. Interview questions can be open-ended and exploratory, and also can be flexible to both interviewers and interviewees (Babbie, 2001, p.258). According to transaction log analysis, some common patterns will be discovered; in terms of forum posts (content) analysis, some behaviors, questions, and challenges will be identified, but the reasons behind them are still unknown. Thus, these patterns/behaviors/challenges are used to design the interview questions, which aim to disclose the grounds of the interacting behaviors. In addition, the proposed initial conceptual model of HDI will also be applied to guide the interview question design. The interview method is a perfect method for deeply understanding user insights, but in the meantime, it could be very expensive (recruiting participants) and time-consuming (implementing interviews and transcribing the recordings). Therefore, the sizes of samples would be greatly limited, in that the generalizability could be questioned. That is why the approaches of transaction log analysis and forum posts (content) analysis are adopted. The three approaches have been individually or collaboratively applied in previous studies related to investigating user behaviors. Combining them can highlight the different strengths to validate and triangulate the findings from each other.

6.3 Implications of H-OGD-I Model

6.3.1 Design Implications

This proposed new H-OGD-I model can assist OGD interface designers in defining advanced and effective user interactions that facilitate users to find, acquire, sense-making, use, and share data.

Design ideas for improving accessibility of OGD. Echoing prior studies that *finding* data is the biggest challenge of interacting with OGD (Lammerhirt, 2017; Zuiderwijk, Janssen, et al., 2012). Our model provides insights on feasible solutions to address the access barriers, such as optimizing the most used channels for accessing data and expanding the least used channel to win more users. Also, we found that the browsing function is more friendly than the keyword search function for beginners. Therefore, a set of clearly defined and easily understood category names are critical for users' effective access experience. In addition, improving OGD platforms' functions of search query recommendations and table search will significantly improve search efficiency.

Moreover, we found that *acquiring* is an essential behavior in the data-seeking process compared to information-seeking behavior, and, yet, according to Study 2, it could be a big challenge when interacting with OGD. For example, most of the difficulties associated with acquiring data are caused by the nonfunctional downloading links. Therefore, a practical suggestion to the designers of OGD portals is to add a regularly automatic checking system for the downloading links. Our results also indicate that a challenging barrier to acquiring data is that users often lack knowledge of APIs, which many portals require for downloading large-sized data. We expect an effective function design that can assist users at all levels (L. Koesten et al., 2021), and the success of obtaining data will immensely influence data use and trust in data and portals.

Design ideas for improving understandability of OGD. The dissertation studies highlight the significance of *making sense* of data. Most OGD is structured data, which requires context to help understand the data. Currently, most datasets are published with a data dictionary explaining the meaning of the fields. However, a more comprehensive contextual

information document is indispensable but deficient in platforms (J. Crusoe et al., 2019; L. M. Koesten, Kacprzak, et al., 2019; Xiao et al., 2019). We suggest that OGD portals collect and publish the critical context information for the datasets. In addition, Study 3 confirms a good visualization can also effectively help users understand data (Ansari et al., 2022; Graves & Hendler, 2013, 2014; Valkanova et al., 2015). For example, one participant prefers a third-party visualization tool over the OGD portals because the data and the patterns in the data are more prominent and more accessible for understanding. More efforts might be needed for OGD portals to develop more effective visualization and interactive tools to assist users with better accessibility, understandability, and usability of the data.

Design ideas for promoting OGD use. A unique behavior of *sharing* data between users who are strangers to each other was revealed in Study 2. Some users are very generous in sharing the data they collected and organized with the public. For instance, “*I have also been compiling pos/neg by zip but only since my post here. I am running a simple cron job to pick it up daily. Wondering if it could be helpful to make it available via this channel or another publicly accessible site, e.g., GitHub.*” Therefore, providing a channel or platform for users to communicate and share data will significantly enhance data use.

6.3.2 Theoretical Implications

This new H-OGD-I model accounts for the stages and user behaviors when interacting with OGD, and therefore it advances the understanding of the complexity of OGD user behaviors. This model is compatible with the existing human data/information interaction models, including J. Crusoe & Ahlin (2019); L. Koesten et al. (2017); Kuhlthau (1991), and the newly discovered behaviors in the stages of *data sensemaking* and *data sharing* make this H-OGD-I model more comprehensive. This dissertation aims to contribute a comprehensive HDI model that is broad enough to describe OGD users’ online interactive behaviors with different tasks and to be applied to various domains in the HDI area. For future research expansions, researchers may apply this model to study other data user groups, such as research data or semi-structured data. Faculty members or librarians may use this model as guidance to structure training for students or people learning how to prepare a task, find,

evaluate, make sense of, and use data to determine how researchers should be led through the process. We expect this model to enrich the theoretical studies in the HDI field.

Although this model was devised for OGD, it can also be generalized to other areas dealing with structured data, especially structured research data. OGD and structured research data possess the same characteristic—structured data, which refers to the data being stored in a standard format and having a standard infrastructure. Therefore, even though the content of government data and research data is different, the logic and infrastructure of data are similar, which causes the behaviors of interacting data to be very similar. In addition, based on the analyses of the forum posts and interview data, we observed that many researchers had used OGD to conduct research studies. Given the similar characteristics of structured data and the overlapped users and task purposes (research), we conclude that this H-OGD-I model may inspire researchers in the field of research data management to understand their user behaviors.

This proposed H-OGD-I model can also be used as a theoretical guide to develop OGD literacy capabilities by researchers librarians and instructors. In fact, Study 5 in this dissertation adopted this new H-OGD-I model to develop the coding schema to guide the analysis of OGD literacy capabilities.

6.4 Implications of the Taxonomy of OGD Literacy Capabilities

6.4.1 Design Implications

This Taxonomy consists of *Action-specific OGD Literacy Competencies Capabilities* and *Action-holistic OGD Literacy Competencies*, which were explored and broken down to each behavioral process when interacting with OGD. This taxonomy was developed purely for user side, excluding data publishers or data professionals who are in charge of data management. Therefore, the abilities to manage data or store data were proposed by other studies (Mandinach & Gummer, 2013; Prado & Marzal, 2013) not contained by it.

Design guide for educational materials. OGD portals can offer interactive tools to facil-

itate users to manipulate data, but we cannot ignore the critical fact that there is a close tie between successfully using data and users' OGD literacy. Based on Study 2, we found that many participants experienced challenges in using data, such as a lack of computational skills or domain knowledge. Additionally, recent studies address the issue of data inequality that has become more prominent with the explosive growth of OGD. There is "a growing gap between those who can work effectively with data and those who cannot" D'Ignazio (2017). Therefore, improving users' OGD literacy is key to promoting data use and data equality. This comprehensive taxonomy can provide a fundamental basis for developing educational materials, including curriculum and workshops. In the two focus groups, all the faculty members and OGD portal administrators acknowledged that this taxonomy will be extremely helpful in facilitating OGD literacy educational project design.

6.4.2 Theoretical Implications

Three research extensions were considered by using this taxonomy. Firstly, this OGD literacy taxonomy can be used to identify research contributions, such as new approaches to specific areas, or to identify behaviors in case studies. To better assist novice users, this taxonomy can help researchers understand the most effective way to visualize data or create interactive tools to lower the cognitive barrier of using data.

Secondly, this taxonomy can guide faculty members to understand the challenges of their students and can also be used as a conceptual scaffolding for the students. For instance, the taxonomy can be designed as a heat map to understand certain issues that occur. A faculty member who teaches a data science class can plot all student issues along this taxonomy; it might demonstrate something about the curriculum, the data sources, or the tools that are being used, and it could be a diagnostic research instrument.

Thirdly, this taxonomy is the beginning of the journey for exploring and identifying OGD literacy capabilities. For future research, researchers can break down the capabilities into different levels of OGD literacy skills to study the measurement of a user's OGD literacy level. As Twidale et al. (2013) argued, to achieve the goals of transparency and participation in open government, enabling all people to have some level of DL is indispensable. However,

the most important point is “some level of DL”— how to define “some level.” Based on the literature review, most research studies discuss how to investigate data literacy capabilities, but very few explore how to evaluate whether an individual is a novice, intermediate, or expert user. For example, if an individual user can make a table, but cannot make visualizations based on the same dataset, is this individual data literate or not? Or, a person can read news with tables or graphs and understand the issues that news delivered. However, if this person cannot make a table or graph, can we say this person is not data literate?

6.5 Design Implications of the Forum Data Analysis

This dissertation detected close and continual communications among users and OGD administrators in the online discussion community. Currently, very few OGD portals have an online forum, but our results demonstrate that establishing an online discussion community benefits both users and OGD administrators. Users with high OGD literacy are willing to help answer questions and share data. Their answers enable other users to obtain access to data, so the overall effort and time for accessing and using OGD become less for new users. At the same time, through an online forum, novices can fulfill their desire to learn by engaging with other users on the platform. Expert users can be considered supporters, and novices are information seekers. We noticed that very few research studies argue for employing social media platforms to empower citizens to learn, understand, and share OGD among themselves. The power and benefits of online communities such as Facebook groups or Google groups for sharing knowledge have been extensively studied and discussed among researchers and practitioners (Lampe et al., 2008; Pi et al., 2013), and their impacts on civic engagement in government work also started to emerge (Gunawong, 2015; Lee & Kwak, 2012). Although social media can potentially play a significant role in promoting civic engagement and improving OGD use, a search with keywords— “open government data,” “civic data,” and “open data” on Facebook, Reddit, and Google groups generated few results. Therefore, we would encourage OGD portals to build an OGD online discussion community to actively and positively interplay with users, to eventually support users.

7.0 Conclusion

The ultimate goal of this dissertation is to increase OGD use and thereby promote data equality for the public. From a user-centered perspective, a sequential, mixed-method design was adopted to conduct five empirical sub-studies: Study 1: observing users' OGD accessing behavioral patterns, Study 2: identifying users' individual OGD behaviors, Study 3: exploring users' cascading OGD behaviors, Study 4: examining user challenges in each H-OGD-I stage, and Study 5: investigate fundamental OGD literacy capabilities.

An initial conceptual model for human OGD interaction was first proposed by combining L. Koesten et al. (2017)'s framework of human structured-data interaction and Pirolli & Card (2005)'s sensemaking model. This model was leveraged to guide the first three sub-studies that answered RQ1 to empirically examine OGD user behaviors when interacting with OGD and formulated a model for human OGD interaction; the findings of the three studies validated and refined the initial conceptual model. Study 4 answered RQ2 to contextualize OGD user challenges when interacting with data in each behavioral process. Study 5 answered RQ3 to identify the needed fundamental OGD literacy capabilities and break them down into behavioral processes.

This dissertation is one of the first research studies to explore OGD user behaviors and create a corresponding behavioral model; it is also one of the first to develop a user-centered fine-grained taxonomy for OGD literacy capabilities. This chapter discusses the significance and contributions of this dissertation, the limitations when conducting this dissertation project, and the future research agenda.

7.1 Significance and Contribution

This dissertation study is built upon a user-centered perspective to explore user OGD behavior and identify the needed OGD literacy capabilities, which can potentially contribute to the fields of HDI theory, data literacy, and open data practice.

Significance to HDI fields: With the development of the open data movement, a massive amount of structured raw data have been available online and will be continually published online. On the other hand, the research on HDI is still an emerging field (Mortier et al., 2013), the studies on HDI are few, and the corresponding models are even fewer. Based on three empirical studies, this dissertation developed a new comprehensive HDI model within the context of OGD (H-OGD-I model) that elaborates the stages and user behaviors when interacting with OGD, and therefore it advances the understanding of the complexity of OGD user behaviors. In addition, because most OGD is structured data, this model can be generalized and applied in other contexts dealing with structured data, e.g., structured research data. With the extensive data available online, more types of data are emerging and developing. Accordingly, the research examining how humans interact with various kinds of data thrives and adds to the continual growth of HDI studies. This comprehensive H-OGD-I model is expected to contribute to the HDI field and serve as a theoretical foundation and guidance for future research.

Significance to data literacy research: Although existing studies have described the critical competencies of data literacy, the competencies that were discovered cannot be generalized to OGD (Koltay, 2015; Prado & Marzal, 2013). Currently, few research studies discuss the data literacy explicitly focusing on OGD. This dissertation study fills this research gap by empirically exploring the fundamental capabilities required for OGD literacy to facilitate a thorough understanding of a needed taxonomy of OGD literacy capabilities. The taxonomy of OGD literacy capabilities was developed from a user-centered perspective, building upon each process of user OGD behaviors, and broken down to a fine-grained level. This comprehensive and granular taxonomy will contribute to supplementing the research in the field of data literacy.

Significance to promoting OGD use and data equality: The theoretical outcome (H-OGD-I model) and the identified contextualized OGD user challenges will contribute to augmenting the efficient and effective use/reuse of OGD. This model enables OGD designers to profoundly understand their users and then provide effective designs for the platform interfaces and tools that can reduce users' cognitive load of using data. Also, the practical outcome (the taxonomy of OGD literacy capabilities) is expected to be a comprehensive

guideline for OGD administrators, librarians, and educators to develop OGD literacy educational materials to improve OGD users' data literacy. In addition, the three research methods (transaction log analysis, forum post analysis, and interview studies) covered a wide range of OGD users, including the population with lower data literacy skills. Therefore, the two outcomes of this dissertation can contribute to enhancing data literacy skills, especially for OGD users with lower data literacy capabilities. Given the similarity of local-level OGD portals across the country, this OGD literacy capability taxonomy could be disseminated and applied to other local-level OGD portals in the U.S., which could promote data equality for the public. Even though OGD has been developed for over a decade, it is still in its initial stage; more and more OGD will become available to the public. Existing studies found that OGD are largely underused (Kawashita et al., 2022; Quarati & De Martino, 2019), and the full potential of OGD can be reached only when most people have the ability to use the data to understand their neighborhoods, communities, and society. This dissertation project will contribute to this ambitious goal; reaching such levels of data literacy will have a transformative effect on society, which can last for years.

7.2 Limitations

Even though I consider this dissertation study to be well designed and conducted, the limitations are still recognized, especially on the data size and research method. This section discusses the limitations of this study.

One limitation of this dissertation study is the data sources and data size. Regarding the forum post studies, although we examined 600 posts and further annotated 238 posts for Studies 2 and 4, and 415 posts for Study 5, this dissertation used the posts from one online discussion community from OpenDataPhilly, which limits the discussions to only concerning one OGD portal. Despite the similarities of local-level OGD portals allowing the findings from the post data to be generalized to OGD users as a whole, observing more online forums from different portals would enable us to further compare the findings to make the exploratory study more convincing. When more active OGD online forums emerge, this

comparison study will become possible in future work. In addition, regarding the interview studies, around half of the interview participants are students, and their experience using OGD is limited to completing the assignments. Though we observed that students' assignments require the users to experience more interactions with data, which strengthens the comprehensiveness of our model, we recommend that future research be expanded to cover more experience in the work environment or real daily life. Additionally, the population of the interview study may be limited. However, the transaction log data from three OGD portals and forum post data with 182 individual users cover a broad group of users. Therefore, this dissertation represents a large portion of the OGD user population, from users with high OGD literacy to those with low OGD literacy.

Another limitation is related to the interview method design. This dissertation adopted three research methods, including quantitative and qualitative, to triangulate and offset the limitations of each method. Nonetheless, there is still space that can be improved, especially for the interview study design. The critical incident technique (CIT) is utilized as the basis for designing the semi-structured interview instrument. The advantage of this technique is that it allows us to understand actual tasks and restore natural processes by asking the participants to recall an entire process of carrying out a real task. However, one limitation of this method is that there could be inconsistencies between what participants described or thought and their real behaviors (Chi et al., 2020). For example, as discussed above, participants did not mention they encountered any difficulties in selecting appropriate datasets for their task, which conflicts with another controlled observation user study (Li et al., 2022) that the evaluation barriers existed in their participants. It is possible that our participants are good at it; therefore, no challenges were brought up, but another experimental user study design could convincingly confirm this assumption.

7.3 Future Work

In the future, my work will focus on continuing to make efforts to improve the use of OGD and promote OGD literacy. This dissertation's findings have provided a clear indication

of the need to increase both the usage and literacy of OGD. To this end, I plan to focus on augmenting the use of OGD through the provision of social support, and enhancing OGD users' literacy by creating an OGD literacy assessment tool. This assessment tool will help to gauge the understanding of OGD users and allow for more effective use of OGD. Furthermore, this tool will be used to measure the progress of OGD users and identify areas of improvement. Ultimately, these efforts will help to foster a better understanding of OGD and its potential applications. This section elaborates on this plan.

7.3.1 Improving OGD Use

Enhance the confidence and trust in OGD by designing a trust rating system.

OGD is becoming increasingly important as we seek to increase trust in governments. The ultimate goal of winning citizens' trust in governments can be accomplished by fostering trust in OGD. By doing so, we can strengthen the trust between citizens and governments. However, previous research studies have shed light on the factors associated with the citizens' trust in OGD, which suggest that citizens do not easily trust the data. For instance, Purwanto et al. (2020) and his team empirically investigated the attributes that can affect citizens' confidence in OGD. They found that system quality and service quality can influence citizens' perspectives of OGD. Another study proposed that when the data is open to the public and the data results can be reused by being replicated, it can contribute to the trust of citizens (Meijer et al., 2014).

Despite the importance of OGD, there are very few studies that provide solutions for increasing the confidence of citizens in it. In order to address this problem, I plan to propose research that aims to cultivate trust for OGD by providing design guidance for a trust rating system. This system would allow users to rate the trust levels for the data that they have used and to provide pertinent comments. The idea is to provide a simple system to allow users to rate the trustworthiness of OGD, as well as to give comments and feedback.

Improving civic engagement through OGD by adopting social media. As argued in the Design Implications section of the Forum Data Analysis (section 6.5), OGD users indeed benefit from the online forum receiving help from others. However, there are very few

active OGD online forums existing, and we also found that though various techniques have been introduced to resolve OGD user difficulties (Kunze & Auer, 2013; Thomas et al., 2015), very few proposal has been presented to employ social media platforms for empowering citizens to learn, share and understand OGD among themselves. As discussed, the power and benefits of online communities such as Facebook groups and Google groups for sharing knowledge on civic engagement in government work have started to emerge (Gunawong, 2015; Lee & Kwak, 2012). Therefore, I attempt to 1) investigate the attitudes, challenges, and needs of OGD portals to apply social media as a connecting platform to improve civic engagement through OGD, 2) examine OGD users' experience, perspectives, behaviors, challenges, and needs of using social media as part of the solution to access and use OGD, and 3) design a system prototype to evaluate the design guidelines developed based on the findings from the first two objectives. This research is expected to 1) encourage more citizens to learn and use OGD to ultimately improve civic engagement and data use, and 2) provide valuable insights into both social media and OGD fields. Applying social media to engage citizens with OGD is new to both social media platforms and OGD portals. Therefore, the findings from this study will benefit both entities in carrying out their goals.

7.3.2 Promoting Data Equality.

Developing the taxonomy for OGD literacy capabilities in this dissertation is the first stage of enhancing OGD users' data literacy. The plan for the next stage is to develop an OGD literacy assessment tool for evaluating users' OGD literacy skills. According to the literature review, I noticed that very few studies were designed to test users' OGD literacy, which could have two explanations. The first possibility is that OGD literacy involves multiple disciplines, such as civic domain knowledge (various domains), statistics, and computational methodology, which can make it difficult to define the evaluation measurements. The second reason is that some competencies cannot be evaluated quantitatively, such as the ability to determine the data needs. This could make it challenging to measure OGD literacy. After identifying the capabilities of OGD literacy in this dissertation, I intend to further investigate the criteria for measuring the capabilities and create an assessment tool to evaluate

users' OGD literacy. The assessment tool will consist of five sections of knowledge and skills: OGD domain knowledge, information literacy, computational thinking, computational skills, and data communication. The test results of the assessment will provide an overall score and scores in each of the five dimensions, allowing users to better understand their OGD literacy capabilities. Furthermore, this test tool will provide personalized learning resource recommendations based on users' testing results.

The first stage of this dissertation project was successful in establishing a comprehensive OGD literacy taxonomy which can be used to design workshops and curricula, as well as improve the interface design of OGD portal platforms to encourage more users to use data. The second stage of the project, the OGD literacy assessment tool, will help to identify and fill in the gaps between the required OGD literacy and users' actual capabilities. Our hope is that this taxonomy and assessment tool will help to reduce data inequality and contribute to the goal of making data more accessible to all.

Appendix A Interview Participant Recruitment Email

My name is Fanghui Xiao, and I am a Ph.D. Candidate at School of Computing and Information, University of Pittsburgh. I am working on a project that is to investigate what specific open government data (OGD) literacy skills are needed when users interact with OGD. We aim to explore this from the perspectives of users. Given that data literacy is a significant factor of using OGD, we have designed the semi-structured interviews to explore user behaviors of interacting with OGD and the needed data literacy skills during the processes. The findings can inform OGD initiatives or librarians to design training guidance and can help OGD designers or developers devise their system and interface more straightforward to use by considering the deficiencies of user data literacy skills. To achieve the goal, we will have interviews mainly focusing on open government data users. Every participant must be 18 years of age or older.

We invite you to participate in an interview study. If you are willing to participate, we will ask about your background (e.g., occupation and primary discipline), your experiences with interacting with OGD data, and the data literacy skills that were used in your previous project. The estimated completion time will be 50 to 60 minutes. You will get a \$20 Amazon card as compensation for your participation.

Your responses will not be identifiable in any way. All answers will be only used for research purposes. Your identity will be kept confidential, and the answers will be revealed only after an anonymous process. Any personal information that could identify you will be removed or adjusted before results are revealed in any way, including publishing, sharing with other researchers, or making datasets public. Your participation is voluntary, and you may withdraw from this project at any time. At last, if you are a Pitt student, participation will not influence your academic standing. Study participation will not negatively or positively impact your grade or course obligations. Please note that audio recording will be taken place for only transcribing purposes. Recordings will not be made to public. You can request to turn off the recording any time. The results will be stored in password-protected computers. The only possible risk is a possible breach of confidentiality. The interviews will be virtually

conducted by Zoom. You have the right to know that your participation is voluntary, and you may stop completing the interview at any time.

This study is being conducted by Fanghui Xiao, who can be reached at fax2@pitt.edu, if you have any questions. You may also contact my advisor Daqing He, PhD. via email: dah44@pitt.edu

If you are willing to help or have any questions, please contact me via email. We very much appreciate your help in advance.

Sincerely,

Fanghui Xiao

Ph.D. Candidate

School of Computing and Information

University of Pittsburgh

Appendix B Interview Protocol

Part 1: Background Information

Q1. Which one of the following best describes your primary discipline?

- ☐ Political Science
- ☐ Economics
- ☐ Sociology
- ☐ Environmental Science
- ☐ History
- ☐ Public Health
- ☐ Library and Information Science
- ☐ Computer Science
- ☐ Information Science
- ☐ Other, please briefly specify:

Q2. Your age group

- ☐ 18-24
- ☐ 25-34
- ☐ 35-44
- ☐ 45-54
- ☐ 55-64
- ☐ 65 and over
- ☐ Prefer not to answer

Q3. How many projects/tasks/assignments have you worked on by using open government data?

- ☐ 1 project
- ☐ 2 projects
- ☐ 3 projects
- ☐ More than 3 projects

Q4. How many times have you used OGD for the projects/tasks/assignments?

- A. 1-5 times (not very familiar)
- B. 6-10 times (familiar)
- C. 15 times above (very familiar)

Part 2: OGD User behaviors

Q5. Could you describe your project? e.g., What is the purpose of your project? And what are the specific objectives of this project?

Q6. What WPRDC/OpenDataPhilly datasets did you use for the project?

Q7. What are the specific processes of interacting with the dataset?

Prompts:

- o look for data first? or have RQs first?
- o I noticed that you use multiple sources. How did you find the data except the datasets from WPRDC Processes:
 - o When finding the data, did you browse the data by categories or search by using keywords. Why did you choose to browse or keyword search?
 - o how do you evaluate if the data is you want? based on what?
 - o Was there a point that you think you find the needed data?
 - o When did you decide to download the datasets?
 - o Were you downloading the data first? Or reading the metadata or data dictionary first?
 - o After downloading the datasets to your local computer, did you go back to WPRDC's website? If yes, for what?
 - o How did you understand the data?
 - o How did you use the data?
 - * clean data
 - * merging data
 - * analyzing data
 - * making decision based on the results

Part 3: OGD User Challenges

Q8. What kinds of challenges do you have when interacting with the datasets that in the project you mentioned before?

Part 4: OGD Literacy

Q9. What data literacy skills are needed when you working on this project? From the starting point to the end.

Appendix C The Model for Human OGD Interaction (H-OGD-I)

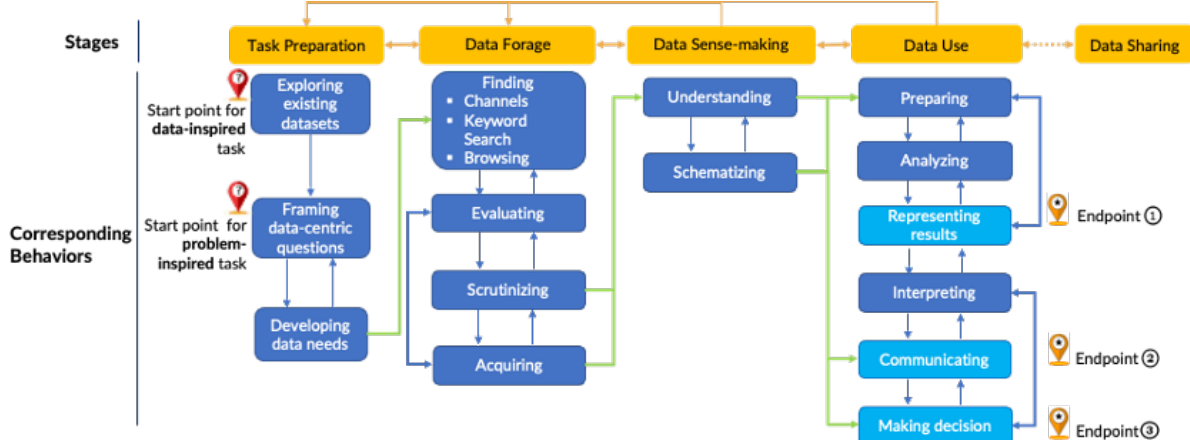


Figure 42: The Model for Human OGD Interaction (H-OGD-I)

Task Types. A *problem-inspired* task refers to a task that starts with the aim of addressing a real problem, then translates the real abstract problem to data-centric questions, to finally address the questions by using data. The behavior of *framing data-centric questions* is the **starting point** of problem-inspired task.

A *data-inspired* task refers to a task that starts from data, through browsing existing data to come up with a real problem, and then translates the real abstract problem to data-centric questions, to finally address the questions by using data. The behavior of *exploring existing datasets* is the **starting point** of data-inspired task.

Task Preparation Stage denotes the process of proposing a real problem that a task aims to address, translating the real problem into data-centric questions, and discovering the data needs for answering the questions. In this study, two types of tasks are identified: question-inspired tasks and data-inspired tasks.

Exploring existing datasets denotes the process when users try to come up with a real problem that is of interest to them by using the existing datasets. The typical behaviors include browsing the categories or tags of datasets and cursorily assessing the topic's relevance and utility of the datasets. Exploring existing datasets is the start point for a data-inspired

task.

Framing data-centric question (s) refers to translating an abstract problem into data-centric questions to address the goal of a task. Framing research questions is the start point for a question-inspired task.

Developing data needs denotes probing into what data are needed to answer the questions proposed by the task, then developing a clear list of required data.

Data Forage Stage denotes the process when users seek data, including the behaviors of finding, evaluating, scrutinizing, and acquiring data. By the end of this stage, users have found relevant and usable datasets.

Finding denotes the process when users look for data, including identifying data sources and applying data-seeking strategies to find data. The primary strategies could include keyword searching and browsing, and or the combination of keyword searching and browsing behaviors.

Evaluating denotes the first look when users evaluate the topic relevance of the identified dataset with their data needs. The actions could involve examining the title and description of the dataset.

Scrutinizing denotes the second look when users rigorously scrutinize the utility of the identified dataset, including examining the file formats, specific data types, data time, missing data, and the accuracy of the data, assessing the variables that can be used to combine multiple datasets.

Acquiring denotes the process when users obtain the datasets. The behaviors may involve clicking the provided link to download the datasets directly and querying data through APIs. The acquiring behavior could happen before or after scrutinizing the data, depending on the users' preference for the data preview function or whether the preview function working well or not.

Data Sensemaking Stage denotes the process when users try to make sense of data, including the behaviors of understanding and schematizing data. By the end of this stage, users have understood the identified data, (e.g., the meaning of column headings and the relationships/logic between data).

Understanding denotes the process when users start to perceive the intended meaning of

the data, such as understanding the meaning of the column headings (field names) and the values of the data. Typical behaviors include closely reading the data dictionaries and/or “read me” files or looking for related concepts from other websites.

Schematizing denotes the process when users try to establish the logical relationships between the data. For example, organizing the data with relevant information or building connections between the data to schematize a simple schematic form of understanding. Typical behaviors include cottoning on the content of data and/or the visualizations provided by platforms, and manipulating data, such as sorting or filtering to find connections between data within one dataset or multiple datasets.

Data Use Stage denotes the process when people apply the collected data to undertake their tasks. The possible activities include preparing data (e.g., cleaning, transforming, and merging), analyzing data, representing results, interpreting and communicating with data, as well as making decisions. This stage could be the last stage of the interacting behaviors, and the outcome of this stage is the answers to the data-centric questions for this task.

Preparing refers to the process of making the data ready to be further analyzed, visualized, or directly used (e.g., building a database). The specific behaviors involve cleaning, extracting, importing, transforming, merging, and reshaping data.

Analyzing denotes the process of utilizing various methods (mainly quantitative) to analyze data to discover useful information, identify patterns, draw conclusions, and/or support decision-making.

Representing results depicts the process of representing the results through different methods. According to our study, the primary methods are visualization and statistical analysis results.

Interpreting indicates the process of exploring and translating the analysis results through combining the comprehension of the pertinent context information and domain knowledge to finally draw conclusions or make decisions. This process translates the data-centric questions back to knowledge or information.

Communicating manifests the process of using appropriate language and manners to communicate the data with the targeted audience, such as telling an evidence-based data story. Making decisions describes the process of making a choice or deciding to take action

based on the understanding when making sense of the identified data, or the interpretations of the analysis results.

End point 1 (Representing results) is the ending point of the task that aims to present the results from OGD interaction. For example, collecting needed data for creating a database. After going through stages of taking preparing, data foraging, and data sense-making, the user prepares the data based on pre-defined requirements and then creates a database to represent the results. The other processes in the data use stage are needless.

End point 2 (Communicating) is the ending point of the task that aims to generate a report for someone else; they do not need to make decisions by themselves, for example, student assignments or work responsibilities.

End point 3 (Making decision) is the ending point of the task that aims to make a final decision by the users themselves after interacting with OGD. The process can be going through all the stages or in the middle of some behaviors (shown in the figure). For example, a user decides to buy a house in certain neighborhoods after schematizing a crime data visualization in an OGD platform (Schematizing to Making decision).

Appendix D The Taxonomy for OGD Literacy Capabilities

Table 17: The Fundamental Action-specific OGD Literacy Capabilities

Behavioral process	Actions	OGD literacy capabilities	Codes
Explore existing datasets	Browse Datasets	Awareness of the existing data sources	ASC-1
Explore existing datasets	Browse Datasets	Ability to follow the infrastructure of the datasets, e.g., the classification structure	ASC-2
Explore existing datasets	Browse Datasets	Ability to browse dataset subjects/categories provided by the data portal to obtain inspiration for initiating a project	ASC-3
Frame data-centric questions	Translate a real problem to data-centric questions	Ability to translate an information-based problem into data-centric questions	ASC-4
Frame data-centric questions	Translate a real problem to data-centric questions	Ability to break a complex problem down into smaller, more manageable questions	ASC-5
Frame data-centric questions	Translate a real problem to data-centric questions	Ability to accurately identify the right questions	ASC-6
Develop data needs	Frame Data Needs	Ability to determine what data is needed to complete the task	ASC-7
Find	Find Sources	Awareness of the existence of data and/or knows where to find the needed data	ASC-8
Find	Find Sources	Ability to identify and locate multiple datasets from different sources to meet the data needs	ASC-9
Find	Search with Keywords	Ability to choose keywords or various combinations of keywords to find the needed data	ASC-10
Find	Browse Data Subjects	Ability to follow conceptualize the infrastructure of the datasets, e.g., the classification structure	ASC-11
Find	Browse Data Subjects	Ability to leverage the provided categories/groups/tags to look for data (knows data categories)	ASC-12
Find	Combine Search Techniques	Ability to combine search methods (e.g., keyword search + subject browse or keyword search + cited-reference search) to seek data	ASC-13
Scrutinize	Scrutinize Usability	Ability to scrutinize if available file formats, data types, and data time period match the needs of the task	ASC-14

Continued on next page

Table 17 – continued from previous page

Behavioral process	Actions	OGD literacy capabilities	Codes
Scrutinize	Scrutinize Usability	Ability to determine if data fields or values are missing	ASC-15
Scrutinize	Scrutinize Usability	Ability to check the accuracy of the data	ASC-16
Scrutinize	Scrutinize Usability	Ability to identify and assess fields that can be used to combine multiple datasets (used in the “Merge” capability)	ASC-17
Acquire	Acquire Data	Ability to download (e.g., through provided links) or collect the needed data manually	ASC-18
Acquire	Acquire Data	Ability to acquire the needed data by using a programming language (e.g. API, SQL)	ASC-19
Understand	Understand Context	Ability to understand when, where, and how the identified datasets were collected, who collected them, and what the dataset is about	ASC-20
Understand	Understand Context	Ability to comprehend this context information, and use it to help understand the data	ASC-21
Understand	Understand Fields	Ability to understand the meaning of column headings and/or the corresponding values by using a provided data dictionary or metadata, or based on personal abilities, such as previous background knowledge, and Google searches	ASC-22
Schematize	Understand Table Structure	Ability to build connections between the columns and/or between tables	ASC-23
Schematize	Understand Table Structure	Ability to understand the relationships between data by manipulating data, such as sorting or filtering	ASC-24
Schematize	Understand Visualizations	Knowledge of the basic visualization categories, such as bar chart, line chart, heat map, and geographic map	ASC-25
Schematize	Understand Visualizations	Ability to interpret the graph’s scale, legend, and axis	ASC-26
Schematize	Understand Visualizations	Ability to identify the data format (geographic data or numerical data) used in the visualization	ASC-27
Schematize	Understand Visualizations	Ability to extract or interpret meaningful information from visualizations using pattern recognition and mathematical thinking	ASC-28
Prepare	Extract Data	Ability to extract the needed granular data (e.g., selecting a subset of rows and columns from a data table)	ASC-29
Prepare	Clean Data	Ability to clean the data through different approaches (identifying anomalies, removing duplicate records, fixing errors, and normalizing values), manually, through a tool (like OpenRefine), or through programming	ASC-30
Prepare	Import Data	Ability to import/upload data to a GIS tool or a programming platform	ASC-31
Prepare	Convert(types or formats)	Knowledge of basic data types: integer, float, string, boolean, character	ASC-32
Prepare	Convert(types or formats)	Knowledge of basic data structure for different data formats (e.g., CSV and JSON)	ASC-33
Prepare	Convert(types or formats)	Ability to convert the data types	ASC-34

Continued on next page

Table 17 – continued from previous page

Behavioral process	Actions	OGD literacy capabilities	Codes
Prepare	Convert(types or formats)	Ability to convert data between different data formats	ASC-35
Prepare	Reshape Data	Ability to aggregate tables of data (e.g., summing columns or creating pivot tables)	ASC-36
Prepare	Reshape Data	Ability to transform data tables to achieve the needed form (e.g., by converting between narrow and wide tables)	ASC-37
Prepare	Merge Data	Knowledge of the basic concept and rules of merging data, such as the need for finding a common field to match on	ASC-38
Prepare	Merge Data	Ability to merge data through different methods, such as tools (Excel) or programming languages	ASC-39
Analyze	Analyze Data	Ability to select the proper analysis method to analyze data in terms of the proposed questions, such as descriptive statistics, exploratory data analysis, and confirmatory data analysis	ASC-40
Analyze	Analyze Data	Ability to analyze data by using statistical/mathematical knowledge and methods (percentage and correlation)	ASC-41
Analyze	Sort Data	Ability to sort data according to the needs	ASC-42
Analyze	Filter Data	Ability to filter data according to the needs	ASC-43
Analyze	Analyze Data	Ability to avoid data analysis errors, e.g., misbeliefs (confusing correlation and causation)	ASC-44
Represent results	Plot Data	Ability to choose appropriate forms of visualizations for the data	ASC-45
Represent results	Plot Data	Ability to create basic visualizations (tables, graphs, and charts) to present data	ASC-46
Represent results	Plot Data	Ability to create geospatial maps to visualize data	ASC-47
Interpret	Distill Meaning	Ability to interpret the patterns or the analysis results	ASC-48
Interpret	Distill Meaning	Ability to draw meaning from results by integrating related contextual information to help make decisions	ASC-49
Communicate	Tell Data Stories	Ability to deliver an informative and easy-to-understand data story, pitch at the correct level for the target audience	ASC-50
Share	Document Data	Ability to describe the processes used to generate the output so that the work can be understood and reproduced (e.g., the descriptions of methodologies for collecting and processing data, commenting code, a laboratory notebook, and limitations of the data)	ASC-51
Share	Document Data	Ability to create data descriptions for data to be shared (e.g., metadata or data dictionary)	ASC-52
Share	Share Data	Ability to upload and edit data on different platforms (e.g., Google Sheets and GitHub)	ASC-53

Table 18: The Fundamental Action-holistic OGD Literacy Competencies

Categories	Actions	OGD Competencies	Codes
Conceptual Foundations	Understand Basic Concept	Ability to understand the basic concepts of data. For example, what is data?	AHC-1
Conceptual Foundations	Understand Basic Concept	Awareness of how essential domain knowledge is to accomplishing a data task (e.g., domain knowledge “reduces the cognitive load” of a data task)	AHC-2
Advanced Actions	Reflect	Ability to reflect on newly obtained feedback/results and apply them to the project practice	AHC-3
Advanced Actions	Reflect	Ability to think back through one’s own work and thought processes, find inconsistencies or opportunities for improvement, and optionally backtrack to repeat an earlier stage in the taxonomy/workflow, making different choices the second time	AHC-4
Advanced Actions	Locate Help	Ability to figure out what help is needed in different processes	AHC-5
Advanced Actions	Locate Help	Ability to find contact information through metadata information or have related background information to locate the contact person for particular datasets	AHC-6
Advanced Actions	Locate Help	Ability to acquire civic domain knowledge through various channels (e.g., workshops, forums, news, and asking related practitioners) and broader background information for datasets, including the origin of a dataset, operational information about how the data was collected, some stories for the dataset, or some information about data publishers	AHC-7
Advanced Actions	Apply Domain Knowledge	Ability to integrate domain knowledge to perform a data task, such as framing data-centric questions, deeply understanding data, and telling a comprehensive data story	AHC-8
Dispositions	Critically Examine Data	Ability to understand the limitations of data. For example, knowing that data is not perfect and data is not absolutely objective	AHC-9
Dispositions	Critically Examine Data	Ability to locate hidden gaps/biases/limitations	AHC-10
Dispositions	Critically Examine Data	Ability to apply a critical lens to examine the biases in the data collection process that may have shaped that data, e.g., the trustability of sources, systemic bias, or selection bias	AHC-11
Dispositions	Operate Ethically	Ability to use data ethically, such as by following related laws, policies, and standards (e.g., the dataset’s license), as well as avoiding disclosing people’s private information	AHC-12
Dispositions	Avoid or Reduce Biases	Ability to reduce non-conscious biases, e.g., confirmation bias and selection bias.	AHC-13
Dispositions	Avoid or Reduce Biases	Ability to avoid misusing or misrepresenting data, e.g., cherry-picking data that supports one’s desired result), that is, demonstrating scientific integrity	AHC-14

References

- Abbas, S., & Ojo, A. (2013). Towards a linked geospatial data infrastructure. In *International conference on electronic government and the information systems perspective* (pp. 196–210).
- Andrejevic, M. (2014). Big data, big questions— the big data divide. *International Journal of Communication*, 8, 17.
- Ansari, B., Barati, M., & Martin, E. G. (2022). Enhancing the usability and usefulness of open government data: A comprehensive review of the state of open government data visualization research. *Government Information Quarterly*, 39(1), 101657.
- Assaf, A., Troncy, R., & Senart, A. (2015). Hdl-towards a harmonized dataset model for open data portals. In *Usewod-profiles@ eswc* (pp. 62–74).
- Audu, A.-R. A., Cuzzocrea, A., Leung, C. K., MacLeod, K. A., Ohin, N. I., & Pulgar-Vidal, N. C. (2019). An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city. In *Conference on complex, intelligent, and software intensive systems* (pp. 224–236).
- Babbie, E. (2001). The practice of social research, wadsworth/thomson learning. *Inc., Belmont, CA*.
- Bates, M. J. (1981). Search techniques. *Annual Review of information Science and Technology*, 16, 139–169.
- Behrens, S. J. (1994). A conceptual analysis and historical overview of information literacy.
- Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications*, 9(3), 379–395.

- Belkin, N. J., Marchetti, P. G., & Cool, C. (1993). Braque: Design of an interface to support user interaction in information retrieval. *Information processing & management*, 29(3), 325–344.
- Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., & Shoup, N. (2015). Beyond data literacy: Reinventing community engagement and empowerment in the age of data.
- Binding, C., & Tudhope, D. (2016). Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries*, 17(1), 5–21.
- Boychuk, M., Cousins, M., Lloyd, A., & MacKeigan, C. (2016). Do we need data literacy? public perceptions regarding canada’s open data initiative. *Dalhousie Journal of Interdisciplinary Management*, 12(1).
- Bruce, H. (2005). Personal, anticipated information need. *Information Research: An International Electronic Journal*, 10(3), n3.
- Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information processing & management*, 31(2), 191–213.
- Cafarella, M. J., Halevy, A., & Madhavan, J. (2011). Structured data on the web. *Communications of the ACM*, 54(2), 72–79.
- Carlson, J., & Johnston, L. (2015). *Data information literacy: Librarians, data, and the education of a new generation of researchers*. Purdue University Press.
- Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, 9(2), 1–22.
- Chatfield, A. T., & Reddick, C. G. (2017). A longitudinal cross-sector analysis of open data portal service capability: The case of australian local governments. *Government information quarterly*, 34(2), 231–243.

- Chi, Y., He, D., Xiao, F., & Zou, N. (2020). Connections and disconnections between online health information seeking and offline consequences. In *Proceedings of the 14th eai international conference on pervasive computing technologies for healthcare* (pp. 73–84).
- Code_of_Federal_Regulations. (2015). Title 48 part 201-299. *US Government Publishing Office, Washington DC*, 593.
- Commission, P. (2017). Data availability and use. Retrieved from: <http://www.pc.gov.au/inquiries/completed/data-access/issues/data-access-issues.pdf>.
- Conradie, P., & Choenni, S. (2014). On the barriers for local government releasing open data. *Government Information Quarterly*, 31, S10–S17.
- Crabtree, A., & Mortier, R. (2015). Human data interaction: historical lessons from social studies and cscw. In *Ecscw 2015: Proceedings of the 14th european conference on computer supported cooperative work, 19-23 september 2015, oslo, norway* (pp. 3–21).
- CraneField, J., Robertson, O., & Oliver, G. (2014). Value in the mash: Exploring the benefits, barriers and enablers of open data apps. In *Proceedings of the european conference on information systems (ecis) 2014, tel aviv, israel, june 9–11, 2014, isbn 978–0–9915567–0–0*.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Crusoe, D. (2016). Data literacy defined pro populo: To read this article, please provide a little information. *The Journal of Community Informatics*, 12(3).
- Crusoe, J., & Ahlin, K. (2019). Users’ activities for using open government data—a process framework. *Transforming Government: People, Process and Policy*.
- Crusoe, J., Simonofski, A., Clarinval, A., & Gebka, E. (2019). The impact of impediments on open government data use: Insights from users. In *2019 13th international conference on research challenges in information science (rcis)* (pp. 1–12).

- Cui, L., & Lee, D. (2022). Ketch: Knowledge graph enhanced thread recommendation in healthcare forums. In *Proceedings of the 45th international acm sigir conference on research and development in information retrieval* (pp. 492–501).
- Dalkey, N., & Helmer, O. (1963). An experimental application of the delphi method to the use of experts. *Management science*, 9(3), 458–467.
- Dawes, S. S., Vidiyasa, L., & Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, 33(1), 15–27.
- Degbelo, A. (2022). Fair geovisualizations: definitions, challenges, and the road ahead. *International Journal of Geographical Information Science*, 36(6), 1059–1099.
- Dervin, B. (1999). Chaos, order and sense-making: A proposed theory for information design. *Information design*, 35–57.
- Dervin, B. (2005). What methodology does to theory: Sense-making methodology as exemplar (chapter 2). In K.E. Fisher, S. Erdelez and L. McKechnie (Eds.). *Theories of Information Behavior*.
- Dervin, B., Clark, K. D., Coco, A., Foreman-Wernet, L., Rajendram, C. P., & Reinhard, C. D. (2012). Sense-making as methodology for spirituality theory, praxis, pedagogy and research. In *Spirituality: Theory, praxis and pedagogy* (pp. 83–94). Brill.
- D’Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal*, 23(1), 6–18.
- D’Ignazio, C., & Bhargava, R. (2016). Databasic: Design principles, tools and activities for data literacy learners. *The Journal of Community Informatics*, 12(3).
- Dorobăț, I. C., & Posea, V. (2021). Open data indicator: An accumulative methodology for measuring the quality of open government data. In *2021 13th international conference on electronics, computers and artificial intelligence (ecai)* (pp. 1–4).

- Dumais, S., Jeffries, R., Russell, D. M., Tang, D., & Teevan, J. (2014). Understanding user behavior through log data and analysis. In *Ways of knowing in hci* (pp. 349–372). Springer.
- Eberhardt, A., & Silveira, M. S. (2018). Show me the data! a systematic mapping on open government data visualization. In *Proceedings of the 19th annual international conference on digital government research: Governance in the data age* (pp. 1–10).
- Eisenberg, M. B., & Berkowitz, R. E. (1990). *Information problem solving: The big six skills approach to library & information skills instruction*. ERIC.
- Elmqvist, N. (2011). Embodied human-data interaction. *CHI 2011*, 104–107.
- Ermilov, I., Auer, S., & Stadler, C. (2013). Csv2rdf: User-driven csv to rdf mass conversion framework. In *Proceedings of the isem* (Vol. 13, pp. 04–06).
- Fechner, T., & Kray, C. (2014). Georeferenced open data and augmented interactive geo-visualizations as catalysts for citizen engagement. *JeDEM-eJournal of eDemocracy and Open Government*, 6(1), 14–35.
- Fidel, R. (2012). *Human information interaction: An ecological approach to information behavior*. Mit Press.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological bulletin*, 51(4), 327.
- Ford, N. (1980). Relating ‘information needs’ to learner characteristics in higher education. *Journal of Documentation*.
- Ford, N. (2015). *Introduction to information behaviour*. Facet Publishing.
- Frank, M., Walker, J., Attard, J., & Tygel, A. (2016). Data literacy-what is it and how can we make it happen? *The Journal of Community Informatics*, 12(3).

- Gascó-Hernández, M., Martin, E. G., Reggi, L., Pyo, S., & Luna-Reyes, L. F. (2018). Promoting the use of open government data: Cases of training and engagement. *Government Information Quarterly*, 35(2), 233–242.
- Gonzalez-Zapata, F., & Heeks, R. (2015). The multiple meanings of open government data: Understanding different stakeholders and their perspectives. *Government Information Quarterly*, 32(4), 441–452.
- Graves, A., & Hendler, J. (2013). Visualization tools for open government data. In *Proceedings of the 14th annual international conference on digital government research* (pp. 136–145).
- Graves, A., & Hendler, J. (2014). A study on the use of visualizations for open government data. *Information Polity*, 19(1-2), 73–91.
- Gray, J., Gerlitz, C., & Bounegru, L. (2018). Data infrastructure literacy. *Big Data & Society*, 5(2), 2053951718786316.
- Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, 70(5), 419–432.
- Gruen, N., Houghton, J., & Tooth, R. (2014). *Open for business: How open data can help achieve the g20 growth target*. Lateral Economics.
- Gunawong, P. (2015). Open government and social media: A focus on transparency. *Social science computer review*, 33(5), 587–598.
- Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday*.
- Haddadi, H., Mortier, R., McAuley, D., & Crowcroft, J. (2013). Human-data interaction (no. ucam-cl-tr-837; p. 9). Retrieved from University of Cambridge, Computer Laboratory website: <http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-837.html>.

- Hornung, H., Pereira, R., Baranauskas, M. C. C., & Liu, K. (2015). Challenges for human-data interaction—a semiotic perspective. In *International conference on human-computer interaction* (pp. 37–48).
- Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. *European journal of ePractice*, 12(1), 4–16.
- Hunt, K. (2005). The challenges of integrating data literacy into the curriculum in an undergraduate institution. *IASSIST Quarterly*, 28(2-3), 12–12.
- Ibáñez, L.-D., & Simperl, E. (2022). A comparison of dataset search behaviour of internal versus search engine referred sessions. In *Acm sigir conference on human information interaction and retrieval* (pp. 158–168).
- Jansen, B. J. (2006). Search log analysis: What it is, what’s been done, how to do it. *Library & information science research*, 28(3), 407–432.
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, 29(4), 258–268.
- Juergens, C. (2020). Digital data literacy in an economic world: geo-spatial data literacy aspects. *ISPRS International Journal of Geo-Information*, 9(6), 373.
- Kacprzak, E., Koesten, L. M., Ibáñez, L.-D., Blount, T., Tennison, J., & Simperl, E. (2019). Characterising dataset search—an analysis of search logs and data requests. *Journal of Web Semantics*, 55, 37–55.
- Kacprzak, E., Koesten, L. M., Ibáñez, L.-D., Simperl, E., & Tennison, J. (2017). A query log analysis of dataset search. In *International conference on web engineering* (pp. 429–436).
- Kassen, M. (2013). A promising phenomenon of open data: A case study of the chicago open data project. *Government information quarterly*, 30(4), 508–513.

- Kassen, M. (2018). Open data and its institutional ecosystems: A comparative cross-jurisdictional analysis of open data platforms. *Canadian Public Administration*, 61(1), 109–129.
- Kawashita, I., Baptista, A. A., & Soares, D. (2022). Open government data use in the brazilian states and federal district public administrations. *Data*, 7(1), 5.
- Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.
- Koesten, L., Gregory, K., Groth, P., & Simperl, E. (2021). Talking datasets—understanding data sensemaking behaviours. *International journal of human-computer studies*, 146, 102562.
- Koesten, L., Kacprzak, E., Tennison, J. F., & Simperl, E. (2017). The trials and tribulations of working with structured data: -a study on information seeking behaviour. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 1277–1289).
- Koesten, L. M., Kacprzak, E., Tennison, J. F., & Simperl, E. (2019). Collaborative practices with structured data: Do tools support what users need? In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–14).
- Koesten, L. M., Mayr, P., Groth, P., Simperl, E., & de Rijke, M. (2019). Report on the data: Search’18 workshop-searching data on the web. In *Acm sigir forum* (Vol. 52, pp. 117–124).
- Koltay, T. (2015). Data literacy: in search of a name and identity. *Journal of documentation*.
- Kremer, N., Bangratz, M., Beetz, J., & Förster, A. (2022). Gis-box improving data literacy in spatial disciplines-integrating spatial data modeling, processing and visualization in spatial study programs.

- Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the analytic hierarchy process. *Government Information Quarterly*, 35(1), 13–29.
- Kučera, J., Chlapek, D., & Nečaský, M. (2013). Open government data catalogs: Current approaches and quality perspective. In *International conference on electronic government and the information systems perspective* (pp. 152–166).
- Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science*, 42(5), 361–371.
- Kunze, S. R., & Auer, S. (2013). Dataset retrieval. In *2013 ieee seventh international conference on semantic computing* (pp. 1–8).
- LAM, M. C. L., URQUHART, C., & BRENDA, D. L. (2016). Sense-making/sensemaking. In *Oxford bibliographies in communication*. Oxford University Press.
- Lammerhirt, M. O. . R. M., D. (2017). The state of open government data in 2017. Retrieved from: <https://blog.okfn.org/files/2017/06/FinalreportTheStateofOpenGovernmentDatain2017.pdf>. ■
- Lampe, C., Ellison, N. B., & Steinfield, C. (2008). Changes in use and perception of facebook. In *Proceedings of the 2008 acm conference on computer supported cooperative work* (pp. 721–730).
- Lathrop, D., & Ruma, L. (2010). *Open government: Collaboration, transparency, and participation in practice*. " O'Reilly Media, Inc."
- Lee, G., & Kwak, Y. H. (2012). An open government maturity model for social media-based public engagement. *Government information quarterly*, 29(4), 492–503.

- Li, A. W., Sinnamon, L. S., & Kopak, R. (2022). Exploring learning opportunities for students in open data portal use across data literacy levels. *Information and Learning Sciences*(ahead-of-print).
- Liu, B., & Jagadish, H. (2009). Datalens: making a good first impression. In *Proceedings of the 2009 acm sigmod international conference on management of data* (pp. 1115–1118).
- Lněnička, M., Nikiforova, A., Saxena, S., & Singh, P. (2022). Investigation into the adoption of open government data among students: The behavioural intention-based comparative analysis of three countries. *Aslib Journal of Information Management*.
- Magalhaes, G., & Roseira, C. (2017). Open government data and the private sector: An empirical view on business models and value creation. *Government Information Quarterly*, 101248.
- Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, 42(1), 30–37.
- Manyika, J., Chui, M., Farrell, D., Kuiken, S. V., Groves, P., & Doshi, E. A. (2013). Open data: Unlocking innovation and performance with liquid information. Retrieved from: <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information>.
- Marchionini, G. (1997). *Information seeking in electronic environments* (No. 9). Cambridge university press.
- Martin, C. (2014). Barriers to the open government data agenda: Taking a multi-level perspective. *Policy & Internet*, 6(3), 217–240.
- Martin, E. G., Law, J., Ran, W., Helbig, N., & Birkhead, G. S. (2017). Evaluating the quality and usability of open data for public health research: a systematic review of data offerings on 3 open data platforms. *Journal of Public Health Management and Practice*, 23(4), e5–e13.

- Masip-Bruin, X., Ren, G.-J., Serral-Gracià, R., & Yannuzzi, M. (2013). Unlocking the value of open data with a process-based information platform. In *2013 ieee 15th conference on business informatics* (pp. 331–337).
- Matthews, P. (2016). Data literacy conceptions, community capabilities. *The Journal of Community Informatics*, 12(3).
- Meijer, R., Conradie, P., & Choenni, S. (2014). Reconciling contradictions of open data regarding transparency, privacy, security and trust. *Journal of theoretical and applied electronic commerce research*, 9(3), 32–44.
- Milic, P., Veljkovic, N., & Stoimenov, L. (2018). Comparative analysis of metadata models on e-government open data platforms. *IEEE Transactions on Emerging Topics in Computing*.
- Millette, C., & Hosein, P. (2016). A consumer focused open data platform. In *2016 3rd mec international conference on big data and smart city (icbdsc)* (pp. 1–6).
- Mortier, R., Haddadi, H., Henderson, T., McAuley, D., & Crowcroft, J. (2013). Challenges & opportunities in human-data interaction. *University of Cambridge, Computer Laboratory*.
- Mutambik, I., Almuqrin, A., Lee, J., Zhang, J. Z., Alomran, A., Omar, T., . . . Homadi, A. (2021). Usability of the g7 open government data portals and lessons learned. *Sustainability*, 13(24), 13740.
- Neumaier, S., & Polleres, A. (2019). Enabling spatio-temporal search in open data. *Journal of Web Semantics*, 55, 21–36.
- Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated quality assessment of metadata across open data portals. *Journal of Data and Information Quality (JDIQ)*, 8(1), 1–29.
- Nikiforova, A., & McBride, K. (2021). Open government data portal usability: A user-centred usability analysis of 41 open government data portals. *Telematics and Informatics*, 58, 101539.
- ODI. (2015). Building global interest in data literacy: A dialogue.

- Ohno-Machado, L., Sansone, S.-A., Alter, G., Fore, I., Grethe, J., Xu, H., ... others (2017). Finding useful data across multiple biomedical data repositories using datamed. *Nature genetics*, 49(6), 816–819.
- Okamoto, K. (2017). Introducing open government data. *The Reference Librarian*, 58(2), 111–123.
- Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., & Ojo, A. (2017). Usability evaluation of an open data platform. In *Proceedings of the 18th annual international conference on digital government research* (pp. 495–504).
- Pászto, V., & Zimmermannová, J. (2019). Relation of economic and environmental indicators to the european union emission trading system: a spatial analysis. *GeoScape*, 13(1).
- Pi, S.-M., Chou, C.-H., & Liao, H.-L. (2013). A study of facebook groups members' knowledge sharing. *Computers in Human Behavior*, 29(5), 1971–1979.
- Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis* (Vol. 5, pp. 2–4).
- Pirolli, P., & Russell, D. M. (2011). *Introduction to this special issue on sensemaking*. Taylor & Francis.
- Prado, J. C., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, 63(2), 123–134.
- Purwanto, A., Zuiderwijk, A., & Janssen, M. (2020). Citizens' trust in open government data: a quantitative study about the effects of data quality, system quality and service quality. In *The 21st annual international conference on digital government research* (pp. 310–318).
- Qin, J., & D'Ignazio, J. (2010). Lessons learned from a two-year experience in science data literacy education. In *Proceedings of the 31st annual iatul conference*.

- Quarati, A., & De Martino, M. (2019). Open government data usage: a brief overview. In *Proceedings of the 23rd international database applications & engineering symposium* (pp. 1–8).
- Radl, W., Skopek, J., Komendera, A., Jäger, S., & Mödritscher, F. (2013). And data for all: On the validity and usefulness of open government data. In *Proceedings of the 13th international conference on knowledge management and knowledge technologies* (pp. 1–4).
- Roa, H. N., Loza-Aguirre, E., & Flores, P. (2019). A survey on the problems affecting the development of open government data initiatives. In *2019 sixth international conference on edemocracy & egovernment (icedeg)* (pp. 157–163).
- Rocca, G. B., Castillo-Cara, M., Levano, R. A., Herrera, J. V., & Orozco-Barbosa, L. (2016). Citizen security using machine learning algorithms through open data. In *2016 8th IEEE Latin-American conference on communications (latincom)* (pp. 1–6).
- Ruijter, E., & Meijer, A. (2020). Open government data as an innovation process: Lessons from a living lab experiment. *Public Performance & Management Review*, 43(3), 613–635.
- Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the interact'93 and chi'93 conference on human factors in computing systems* (pp. 269–276).
- Saxena, S. (2018). Drivers and barriers towards re-using open government data (ogd): a case study of open data initiative in oman. *foresight*.
- Shields, M. (2005). Information literacy, statistical literacy, data literacy. *IASSIST quarterly*, 28(2-3), 6–6.
- Spitzer, K. L., Eisenberg, M. B., & Lowe, C. A. (1998). *Information literacy: Essential skills for the information age*. ERIC.

- Stephenson, E., & Caravello, P. S. (2007). Incorporating data literacy into undergraduate information literacy programs in the social sciences: A pilot project. *Reference services review*.
- Swamiraj, M., & Freund, L. (2015). Facilitating the discovery of open government datasets through an exploratory data search interface. In *Open data research symposium*.
- Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3), 178–194.
- Thomas, P., Omari, R., & Rowlands, T. (2015). Towards searching amongst tables. In *Proceedings of the 20th australasian document computing symposium* (pp. 1–4).
- Trajkova, M., Alhakamy, A., Cafaro, F., Mallappa, R., & Kankara, S. R. (2020). Move your body: Engaging museum visitors with human-data interaction. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–13).
- Twidale, M. B., Blake, C., & Gant, J. P. (2013). Towards a data literate citizenry.
- Tygel, A. F. (2016). Semantic tags for open data portals: Metadata enhancements for searchable open data. *Federal University of Rio de Janeiro*.
- Tygel, A. F., Auer, S., Debattista, J., Orlandi, F., & Campos, M. L. M. (2016). Towards cleaning-up open data portals: A metadata reconciliation approach. In *2016 ieee tenth international conference on semantic computing (icsc)* (pp. 71–78).
- Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. (22). doi: <https://doi.org/https://doi.org/10.1787/5k46bj4f03s7-en>
- Valkanova, N., Jorda, S., & Moere, A. V. (2015). Public visualization displays of citizen data: Design, impact and implications. *International Journal of Human-Computer Studies*, 81, 4–16.

- Van Veenstra, A. F., & Van Den Broek, T. A. (2013). Opening moves—drivers, enablers and barriers of open data in a semi-public organization. In *International conference on electronic government* (pp. 50–61).
- Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*, 31(2), 278–290.
- Verburg, P. H., Neumann, K., & Nol, L. (2011). Challenges in using land use and land cover data for global change studies. *Global change biology*, 17(2), 974–989.
- Verdegem, P., & Verleye, G. (2009). User-centered e-government in practice: A comprehensive model for measuring user satisfaction. *Government information quarterly*, 26(3), 487–497.
- Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly*, 33(2), 325–337.
- Victorelli, E. Z., Dos Reis, J. C., Hornung, H., & Prado, A. B. (2020). Understanding human-data interaction: Literature review and recommendations for design. *International Journal of Human-Computer Studies*, 134, 13–32.
- Viera, A. J., Garrett, J. M., et al. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5), 360–363.
- Waal, S. v. d., Wezel, K., Ermilov, I., Janev, V., Milošević, U., & Wainwright, M. (2014). Lifting open data portals to the data web. In *Linked open data—creating knowledge out of interlinked data* (pp. 175–195). Springer.
- Wagenaar, T. C. (2004). Is there a core in sociology? results from a survey. *Teaching Sociology*, 32(1), 1–18.
- Walravens, N., Breuer, J., & Ballon, P. (2014). Open data as a catalyst for the smart city as a local innovation platform. *Communications & Strategies*(96), 15.

- Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., & Dwivedi, Y. K. (2017). Open data and its usability: an empirical view from the citizen's perspective. *Information Systems Frontiers*, 19(2), 285–300.
- Wilson, T. D. (1994). Information needs and uses: fifty years of progress. *Fifty years of information progress: a Journal of Documentation review*, 15–51.
- Wilson, T. D. (1999). Models in information behaviour research. *Journal of documentation*, 55(3), 249–270.
- Wilson, T. D. (2000). Human information behavior. *Informing science*, 3(2), 49–56.
- Wirtz, B. W., Weyerer, J. C., Becker, M., & Müller, W. M. (2022). Open government data: A systematic literature review of empirical research. *Electronic Markets*, 1–24.
- Wolff, A., Gooch, D., Montaner, J. J. C., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*, 12(3).
- Wolff, A., Montaner, J. J. C., & Kortuem, G. (2016). Urban data in the primary classroom: bringing data literacy to the uk curriculum. *The Journal of Community Informatics*, 12(3).
- Xiao, F., He, D., Chi, Y., Jeng, W., & Tomer, C. (2019). Challenges and supports for accessing open government datasets: Data guide for better open data access and uses. In *Proceedings of the 2019 conference on human information interaction and retrieval* (pp. 313–317).
- Xiao, F., Jeng, W., & He, D. (2018). Investigating metadata adoptions for open government data portals in us cities. *Proceedings of the Association for Information Science and Technology*, 55(1), 573–582.
- Xiao, F., Lyon, L., Zou, N., & Gradeck, R. M. (2018). Emerging roles for optimising re-use of open government data. *International Journal of Digital Curation*, 13(1).

- Xiao, F., Thaker, K., & He, D. (2022). Categorizing open government data users by exploring their challenges and proficiency. In *Chi conference on human factors in computing systems extended abstracts* (pp. 1–7).
- Yang, D., Kraut, R. E., Smith, T., Mayfield, E., & Jurafsky, D. (2019). Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–14).
- Zuiderwijk, A., & Janssen, M. (2014). Barriers and development directions for the publication and usage of open data: A socio-technical view. In *Open government* (pp. 115–135). Springer.
- Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., Alibaks, R. S., & Sheikh_Alibaks, R. (2012). Socio-technical impediments of open data. *Electronic Journal of e-Government*, 10(2), 156–172.
- Zuiderwijk, A., Janssen, M., & Susha, I. (2016). Improving the speed and ease of open data use through metadata, interaction mechanisms, and quality indicators. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2), 116–146.
- Zuiderwijk, A., Jeffery, K., & Janssen, M. (2012). The potential of metadata for linked open data and its value for users and publishers. *JeDEM-eJournal of eDemocracy and Open Government*, 4(2), 222–244.