Ming Zhan

# Public Libraries Meet Big Data: Roles, Comprehension and Practical Applications

Åbo Akademi University

# Ming Zhan

Born in 1988

## Degrees

Bachelor in E-commerce 2011

MSc in Corporation Management 2014

MSc in Economics and Business Administration 2016

# Public Libraries Meet Big Data: Roles, Comprehension and Practical Applications

Ming Zhan

Academic Dissertation

to be presented with the permission of
the Faculty of Social Sciences, Business and Economics, Åbo Akademi University,
for public examination Online, Turku, Finland
on Friday, June 18, 2021, at 13 o'clock

Opponent is Professor Mike Thelwall
from University of Wolverhampton

Information Studies
Åbo Akademi University, 2021

**Dedicated to my beloved mother!**

ISBN 978-952-12-4069-0 (printed)
ISBN 978-952-12-4070-6 (digital)
Painosalama, Turku, Finland 2021

# Acknowledgements

I have been dreaming for a long time about what it would be like to start to write my acknowledgements for the PhD thesis. Meanwhile, this is also difficult for me because words cannot convey how much I appreciate the care and help I received from every beloved person in during my doctoral study.

First of all, I would like to thank Åbo Akademi University, Jubileumsfonden and City of Turku for their financial support of my PhD studies.

Thank you, my supervisor Professor Gunilla Widen. I still remember the first time I went to your office to talk about my academic plan: you were so supportive and elegant. You are tolerant and helpful. You reply to all my emails and messages. You encourage me even for small achievements in my work. To choose you as my PhD supervisor is the best decision I have ever made.

Thank you, Professor Hazel Hall. Thanks for being my external supervisor and widening my perspectives. You brought many hopes that have supported me in my PhD work.

Thank you, Professor Haibo Lee. You have faith in me in spite of my less relevant experience in Big Data Analytics. You have put in great effort and provided constructive suggestions on my papers. Your acceptance of me as a guest researcher in your department improved my technical skills to another level.

Thank you, Professor Mike Thelwall for being my opponent and providing useful insights and opinions to improve my research.

Thank you, Professor Nils Pharo and Professor Birger Larsen for your concrete comments on my dissertation.

Thank you, Qin Yu. I must have done something right to deserve you in my life. Meeting you is a pleasant gift in my life. Without your support financially and mentally, I would suffer a lot.

Last but not least, I would also appreciate all the supports and suggestions from Jannica Heinström, Anu Ojaranta, Xiaofei li, Pengcheng Ye, Rui Yang, Fang Liu, Guopeng Yu, Ye Hong, Xun Liu, Jamie Johnston and other friends, colleagues, and reviewers. Without your help, I would not go this far! Words are not enough to represent my sincerely grateful heart. I think one thing that I can do to reward all these great people is to commit myself to help other people in their academic work and personal life: to pass on all the support and love I have ever got. Many thanks to all of you.

## List of Original Publications

1. Zhan, M., & Widén, G. (2018). Public Libraries: Roles in Big Data. The Electronic Library, 36(1), 133–145. https://doi.org/10.1108/EL-06-2016-0134

2. Zhan, M., & Widén, G. (2019). Understanding Big Data in Librarianship. *Journal of Librarianship and Information Science*, *51(2)*, 561–576. https://doi.org/10.1177/0961000617742451

3. Zhan, M., Yu, Q., & Wang, J. (2020). Effectively Organizing Hashtags on Instagram: A Study of Library-related Captions. *Information Research*, *25*(2). http://www.informationr.net/ir/25-2/paper858.html

4. Zhan, M., Tu, R., & Yu, Q. (2018). Understanding Readers: Conducting Sentiment Analysis of Instagram Captions. In *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence* (pp. 33–40). https://dl.acm.org/doi/10.1145/3297156.3297270

All articles are reprinted with permission

## Author Contributions

The author drafted the manuscripts for all four studies included in the dissertation. In Study 1 and Study 2, the author planned and designed the study together with the co-author. Data collection and data analysis were conducted by the author. In Study 3, the author was in charge of the general work, designed the research and analysed the data. Ji Wang worked in an advising role in research design and data analysis. Qin Yu planned and accomplished data collection in Study 3 and Study 4 together with the author. The author and Qin Yu also worked together to maintain the database used in Study 3 and Study 4. In study 4, the author took a sole role in designing the study and in analyzing the research data. Ruibo Tu assisted in the design of the empirical research, and discussed the implementation of various machine learning algorithms.

# Abstract

The world is witnessing the advent of Big Data. Meanwhile, the development of technology accelerates the spread of Big Data. As such, organizations have started to realize that Big Data can be important resource owing to its value transformation characteristic. Both private and public sectors have applied Big Data for resource saving, decision making, service improving etc. Public libraries as one of the necessary parts of public sectors should also commit themselves into exploiting the potential values of Big Data.

Furthermore, the wide application of social media not only brings about more possibilities for public libraries to extend their services, communicate with users, and present themselves, but also makes public libraries confront an exponential data explosion. Since Big Data can be transformed into useful information, public libraries, as hubs of information, are in a natural position to explore values via analyzing Big Data generated on social media, and manage information and knowledge generated from Big Data. Nonetheless, there are few studies focusing on helping public libraries to understand, manage and use Big Data. Therefore, the overall purpose of this study is to help public libraries realize what their responsibilities might be in the context of Big Data and to understand what Big Data is and how it can be applied.

In order to achieve the overall purpose, four research questions are asked. Each question is answered by one study included in this doctoral dissertation. Inductive approaches combining qualitative or quantitative methodologies are conducted to accomplish each study.

Q1: What kinds of roles should public libraries undertake in the context of Big Data? An online survey and eleven semi-structured interviews with library directors were carried out to identify roles of public libraries in the context of Big Data. Q2: What does Big Data mean specifically in librarianship? A content analysis was conducted to highlight key aspects of Big Data definitions used in library and information science literatures. The answers to these two questions jointly lay the theoretical foundation of Big Data for public libraries.

In order to present public libraries with concrete examples about how to apply Big Data, Instagram is chosen as the representative of social media to collect data owing to its rising popularity, its leading role in image-based social media and the lack of research in utilizing Instagram in public libraries. Hashtags are chosen as the starting point to design research projects, owing to their widespread usage on social media, in particular Instagram. Hashtags could signify the content of Instagram captions and boost communication between caption posters and other Instagram users. Therefore, two research questions are asked based on these two functions of hashtags. Q3: How should libraries effectively organize hashtags to attain more “likes” and comments for library-related posts on Instagram? Q4: What

topics do current readers like or dislike? Millions of library-related captions were collected and analysed to answer these two questions via regression models and supervised machine learning models.

In the end, this study outlines nine roles for public libraries to undertake in the context of Big Data. A Big Data definition specifically used in librarianship is also put forward. Two applications of Big Data for public libraries are organized. These three contents together fulfill the overall purpose. The accomplishment of this study fills research gaps in bringing Big Data to public libraries, enriches the content of Big Data applications and Instagram applications in public libraries, handles the uneven spread research in social media study regarding the single-platform prevalence, suggests a novel way to use hashtags: hashtag organization, and provides a creative way to know library users: sentiment analysis on hashtags. Moreover, this doctoral study is organized in Finland where public libraries are highly developed. Therefore, the result of this study could contribute to the development of public libraries in the context of Big Data in other countries.

# Svensk sammanfattning

Världen bevittnar fördelarna med big data, allt medan den tekniska utvecklingen accelererar och producerar stora mängder data. Olika organisationer har börjat inse att big data är en viktig resurs som kan skapa mervärde. Både den privata och den offentliga sektorn har börjat använda sig av big data för att stöda beslutsfattande, utveckla tjänster osv. De allmänna biblioteken, som är en viktig del av den offentliga sektorn, bör också utreda hur de kan utnyttja det potentiella värdet som big data för med sig.

Den utbredda användningen av sociala medier medför inte bara flera möjligheter för allmänna bibliotek att utöka sina tjänster, kommunicera med användare och presentera sig själva, utan innebär också utmaningar med att hantera en exponentiell ökning av data. Eftersom big data kan omvandlas till användbar information, kunde de allmänna biblioteken vara i centrum för denna transformation. De innehar en naturlig position för att utforska användarmönster, t.ex. genom att analysera big data som genereras på sociala medier och därmed hantera information och kunskap som genereras från denna data. Ändå finns det få studier som fokuserar på att stöda de allmänna biblioteken att förstå, hantera och använda big data. Därför är det övergripande syftet med denna studie att bidra med insikter som kan hjälpa allmänna bibliotek med att förstå vad big data är, hur det kan tillämpas och således utveckla de allmänna bibliotekens ansvarsområden i samband med big data.

För att uppnå det övergripande syftet ställs fyra forskningsfrågor. Varje fråga besvaras av en studie som ingår i denna doktorsavhandling. Induktiva metoder som kombinerar kvalitativa eller kvantitativa metoder genomfördes för att genomföra studierna.

1) Vilka roller kan allmänna bibliotek ha i samband med big data? En online surveyundersökning och elva semistrukturerade intervjuer med biblioteksdirektörer genomfördes för att identifiera biblioteks olika roller i samband med big data. 2) Vad betyder big data specifikt i bibliotekskontext? En innehållsanalys av relevant litteratur genomfördes för att lyfta fram centrala big data-definitioner. Svaren på dessa två frågor utgör den teoretiska grunden för hur big data definieras och förstås inom allmänna bibliotek.

För att presentera konkreta exempel på hur man kunde använda big data inom allmänna bibliotek, valdes Instagram att representera sociala medier och som plattform för att samla in data. Valet har gjorts på grund av att Instagram hela tiden ökar i popularitet samt för dess ledande ställning bland bildbaserade sociala medier. Det finns också en brist på forskning om användandet av Instagram data inom allmänna bibliotek. Hashtags valdes som utgångspunkt för att designa studierna, de beskriver bildernas innehåll,

de fungerar som verktyg för kommunikation mellan bilder och användare och användningen av hashtags är mycket utbredd på sociala medier, särskilt Instagram. Två forskningsfrågor ställdes, baserade på två olika hashtagsfunktioner. 3) Hur kan biblioteken effektivt organisera hashtags för att uppnå ökad interaktion med bibliotekrelaterade inlägg på Instagram i former av "gillanden" och kommentarer? 4) Vilka ämnen gillar eller ogillar användarna? Flera miljoner biblioteksrelaterade bildtexter samlades in och analyserades för att svara på dessa två frågor via regressionsmodeller och maskininlärningsmodeller.

Resultaten från studierna bidrog till att definiera nio olika roller som allmänna bibliotek kan ta för att bättre använda big data i sin verksamhet och en definition av big data, som specifikt kan används i bibliotekskontext, presenteras. Två applikationer av big data för allmänna bibliotek utvecklades för denna studie. Dessa resultat uppfyller tillsammans det övergripande syftet och denna studie bidrar således konkret till forskningen om de möjligheter och det mervärde som big data kan föra med sig till allmänna bibliotek. Avhandlingen bidrar även till utvecklandet av big data och Instagram-applikationer i allmänna bibliotek och presenterar nya sätt att använda hashtags, att på ett kreativt sätt bättre lära känna sina biblioteksanvändare, dvs. sentimentanalys av hashtags. Denna studie utfördes i Finland som har ett välfungerande biblioteksystem. Resultaten av denna studie kan förhoppningsvis bidra till utvecklandet av big data vid allmänna bibliotek även i andra länder.

# Table of Contents

# List of Figures and Tables

# 1.Introduction

Since the year 2011, Big Data has won attention in both industries and academia (Kho, 2018; Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015). According to Kho (2018), the rise of Big Data could be attributed to the growth of the Internet of Things (IoT), which is a new technology paradigm outlining a network of machines and devices interacting with each other (Lee & Lee, 2015). The report of the International Telecommunications Union (Peña-López, 2005) states that the IoT represents technical innovations such as ratio-frequency identification, sensor technologies, and embedded intelligence, which lead to the generation of a vast amount of data and require continuous development of tools and techniques to manage a huge amount of data. As such, the breeding ground of Big Data is cultivated. Based on the forecast on Statista (Holst, 2020), the amount of data worldwide in 2025 will be 175 zettabytes, more than three times the data amount in 2020. Simply put, the wave of Big Data is arriving.

Big Data is defined as "the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value." (De Mauro, Greco, & Grimaldi, 2016, p.24). The transformation into value implies that Big Data, wisely processed, has great potential to generate useful information (Rajaraman, 2016) for both private sectors (Banica & Hagiu, 2015; Wamba et al., 2015) and public sectors (Klievink, Romijn, Cunningham, & de Bruijn 2017). Big Data applications show substantial potential in saving resources, providing sustainable services, and attaining high efficiencies (Desouza & Jacob, 2017; Kim, Trimi, & Chung, 2014; Malomo & Sena, 2017).

As one of the necessary parts of the public sector, public libraries are indispensable in exploiting the beneficial applications of Big Data. Furthermore, the wide application of social media makes public libraries confront an exponential data explosion. According to Arnaboldi et al. (2017), one of the side effects of the social media explosion is the advent of Big Data. Nowadays, millions of people are connected on social media, owing to the rapid spread of communication technologies and their ease of use (Arnaboldi et al., 2017; Dalla Valle & Kenett, 2018; de Zuniga & Diehl, 2017). Since Big Data can be transformed into useful information, public libraries can explore user patterns via analyzing Big Data generated on social media to win user attention and improve communication with users.

However, few studies focus on the benefits of social media for public libraries from a Big Data point of view. There are 3.5 billion active users of social media every single day (Ortiz-Ospina, 2019). Their actions on social media, such as commenting, posting, or adding "likes", create a vast amount of data, which indicates that wisely analysed, such data can be transformed into valuable information to understand users. Public libraries, as hubs of information, are in a natural position to carry out this transformation and

manage information and knowledge generated from Big Data. Nonetheless, few studies are focusing on helping public libraries to understand, manage and use Big Data.

Therefore, this doctoral study is conducted to help public libraries understand their responsibilities in the context of Big Data, comprehend Big Data, and attain useful information from Big Data. Social media has been chosen as the platform to collect Big Data. To narrow the scope of the study, Instagram was chosen as the representative of social media. The value of Big Data in this study is reflected by the analysis of hashtags in millions of library-related Instagram captions to enhance the interaction between public libraries and users and to help public libraries understand their users.

## 1.1 Purpose and aims

The overall purpose of this study is to help public libraries realize what their responsibilities might be in the context of Big Data and to understand what Big Data is and how it can be applied. This overall purpose is fulfilled by achieving three subordinate aims:

Aim 1: To identify roles of public libraries in the context of Big Data
Aim 2: To define Big Data specifically in librarianship
Aim 3: To provide practical examples of Big Data applications for public libraries

The first two aims lay the theoretical foundation of Big Data for public libraries. The definition of Big Data, the characteristics of Big Data, and the roles and responsibilities of public libraries are all addressed through these two aims. The third aim provides an opportunity to explore the value of Big Data. Together, these aims help to ensure the fulfilment of the overall purpose of this doctoral study.

In order to accomplish the aims, four studies have been conducted by answering four specific research questions:

Q1: What kinds of roles should public libraries undertake in the context of Big Data?

Q2: What does Big Data mean specifically in librarianship?

Q3: How should libraries effectively organize hashtags to attain more “likes” and comments for library-related posts on Instagram?

Q4: What topics do current readers like or dislike?

Each research question is studied in a separate study as part of the dissertation work. The interaction of these four research questions constitutes the framework of this study, as illustrated in Figure 1:

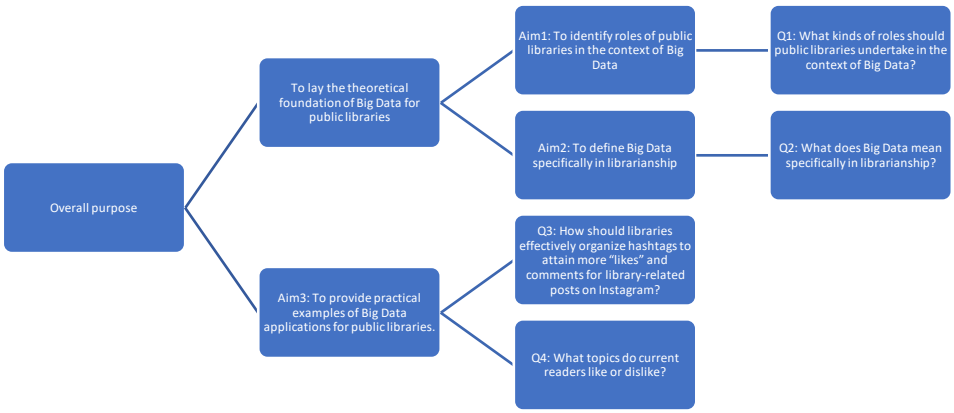



**Figure 1. Research framework**

As shown in Figure 1, Aim 1 and Aim 2 jointly lay the theoretical foundation of Big Data for public libraries. Each of these aims is fulfilled by one research question, which is addressed by a corresponding study.

Study 1: Outlining roles of public libraries in the context of Big Data

This study is designed to help public libraries pinpoint their responsibilities in the wave of Big Data. By knowing their potential roles, public libraries could renew or arrange their daily services to respond to the advent of Big Data.

Study 2: Comprehending Big Data in librarianship

This study aims to answer the question: what does Big Data mean specifically in librarianship? A definition of Big Data is expected so as to help librarians realize what Big Data is in their daily work.

Aim 3 focuses on presenting public libraries with examples about how to apply Big Data to attain useful insights and patterns. Aim 3 is reached via two quantitative studies analysing hashtags on Instagram.

In order to provide concrete examples of Big Data applications, hashtags – words, phrases, and alphanumeric characters preceded by the symbol "#" (Dumbrell & Steele, 2015; Fink, Schmidt, Barash, Cameron, & Macy, 2016) – are chosen as the starting point to design research projects, owing to their widespread usage on social media, in particular Instagram. According to Zappavigna (2015), hashtags can indicate the semantic domain and enlarge the audience. That is to say, wisely used, hashtags could signify the content of Instagram captions and boost communication between caption posters

and other Instagram users. Therefore, the two projects are planned based on these two functions of hashtags.

Study 3: Effectively organizing hashtags on Instagram

This study is conducted so as to answer the research question: how to effectively organize hashtags to attain more "likes" and comments for library-related posts on Instagram? This project emphasizes the communication function of hashtags, highlighting the best location of hashtags in a caption to attract attention (in this research, represented by the number of "likes" and comments) and thus to boost communication with others. Millions of library-related captions are collected and analysed. The findings of this study would be helpful for librarians in communicating with library patrons and marketing library events or resources in a more effective way, which in the end will give them a practical example to apply Big Data in their work.

Study 4: Outlining topics current readers like or dislike

This project focuses on the content identification function of hashtags. Millions of captions with the hashtag "#read" or "#reading" are collected and analysed. Popular hashtags in these captions are outlined, which demonstrate the popular topics among current readers. In addition, the opinion towards these popular topics is analysed so as to pinpoint which topics are liked or disliked by current readers. With the knowledge of reader preference, the public library could strategically arrange their resources to understand their readers better and thus to improve their services. This study is another example for public libraries to understand the power of Big Data applications.

All in all, the whole research can provide insights for public libraries that intend to involve Big Data and Instagram. With the accomplishment of these four studies, the overall aim could be reached.

## 1.2 Motivation and contribution

### 1.2.1 Motivation

The motivation for this study is explained by answering the following three questions:

1) Why Big Data in the public library?

In academia, research on Big Data and the public library is in its infancy, as is presented by Figure 2 and Figure 3:

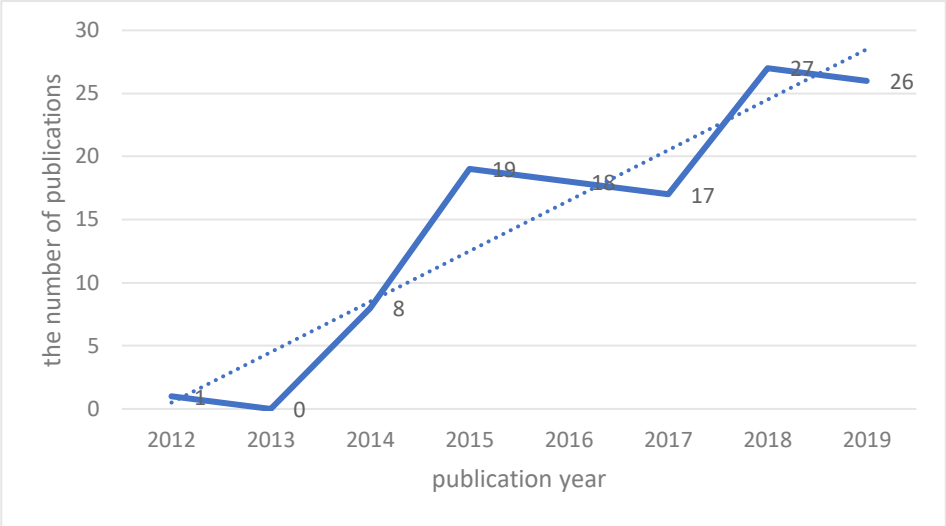



**Figure 2. Publication distribution in years**

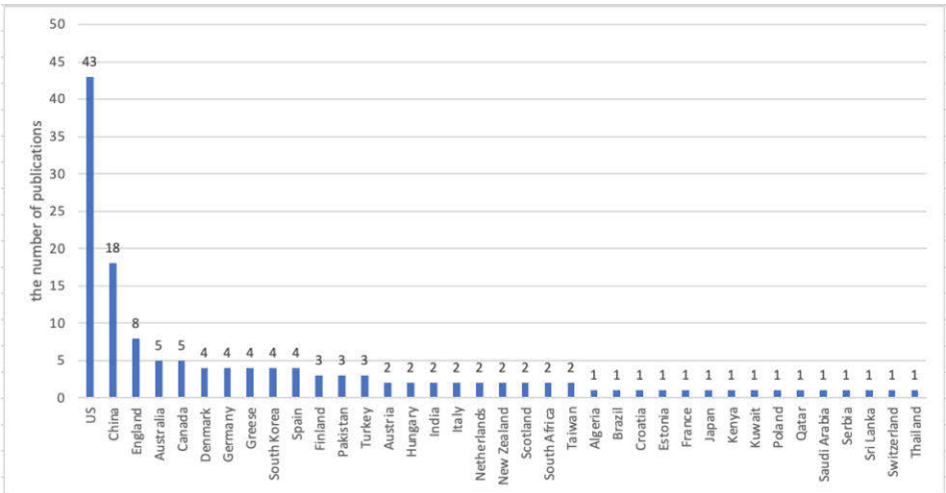



**Figure 3. Publication distribution by country**

The result was retrieved on 17 June 2020, searching Library and Information Science on Web of Science, with "Librar*" and "Big Data" in topic (language: English).

According to Figure 2, the starting year of studies in Big Data and libraries is 2012. From this year onwards, there is an increasing trend to study Big Data in a library context. This indicates that more and more scholars in the field of library and information science (LIS) have paid attention to aspects of Big Data. However, the number of studies in a single year remains fewer than 30, which implies a current lack of attention in this field. Furthermore, Figure 3 shows that scholars in the USA and China are pioneers in this field, while most other countries have fewer than five total publications as of June 2020. As for Finland, mere three studies have been published as of 2019, which demonstrates a significant lack of attention in Finland compared with

the pioneer countries. Therefore, this study is motivated to fill this research gap and make a contribution to Finnish librarianship.

In addition, the author had private talks with librarians working in Finnish public libraries while conducting the academic project Big Cities meet Big Data (Widén et al., 2016). During these conversations, librarians kept mentioning that they were keen to know how to use Big Data in the context of public libraries. This doctoral study is encouraged by this curiosity, and it aims at combining Big Data with the public library. Moreover, all these librarians could foresee the significance of managing Big Data in the public library. However, they currently have no concrete solutions for involving Big Data in their daily responsibilities. This indicates that there is a gap in utilizing Big Data in the public library, at least in Finnish public libraries.

2) Why Big Data from social media?

Under the newly enacted Finnish Library Act from 2017 (Finnish Ministry of Education and Culture, 2017), the objectives of Finnish public libraries are: (1) providing library and information services for citizens to pursue self-cultivation and learning, (2) establishing virtual and interactive networks, and (3) encouraging active citizens, democracy, and freedom of speech. Interactions between Finnish public libraries and the citizens are emphasized. Sandelius (2012) stated that Finnish libraries are keen to know citizens and look for new media to maintain a high standard of services in the digital revolution. Such objectives and expectations make social media a suitable platform to pursue this study in the context of Finnish public libraries. First of all, user-generated social media content (e.g., comments, captions, and "likes") could represent their interaction with public libraries. Wisely analysed, patterns could be identified to boost the interaction. Secondly, Fan & Gordon (2014) confirm that social media activities are regular for adults. Thus, their opinions and trends and the social networks among them can be gleaned with suitable analytic methods, making inroads for library services in order to better interact with library users. As such, focusing on Big Data generated on social media can not only fulfill the research aim but also conform with the Finnish Library Act.

3) Why Instagram?

The overall movement towards image-based platforms and the disproportionate coverage of studies on social media are the main reasons that this doctoral study chooses Instagram as the representative of social media.

First of all, the popularity of Instagram is increasing, as is shown in Figure 4.

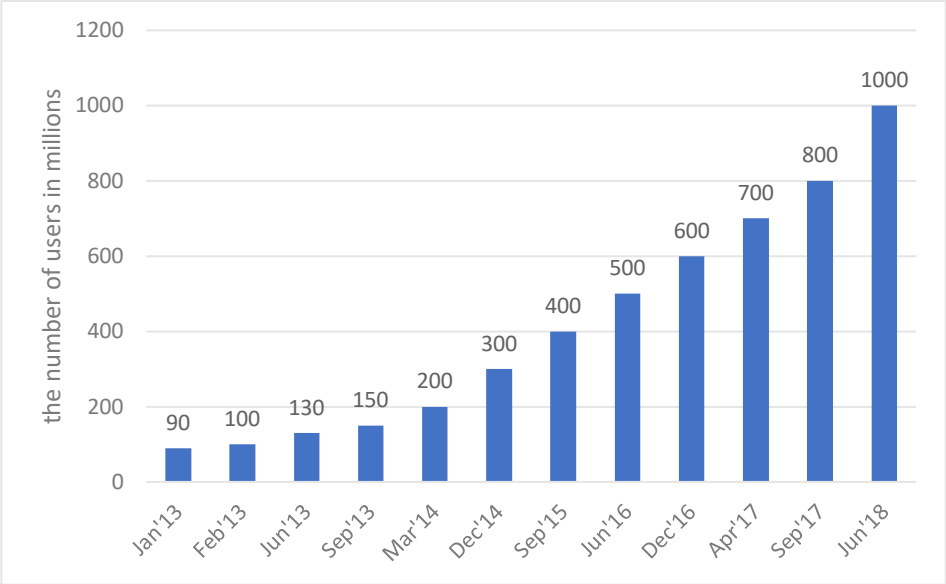



**Figure 4. The number of active Instagram users in millions**

(Clement, 2019)

The number of active Instagram users is dramatically increasing. This trend indicates that more people are using Instagram and thus more data is being generated on Instagram, which justifies Instagram as a solid platform for collecting Big Data.

Secondly, according to the study by Stuart, Stuart, & Thelwall (2017), social media has evolved from text-intensive services (such as wikis and blogs) to message-based services (Twitter) to social networking sites (Facebook) to multi-media services (YouTube). The current pioneering social media is image-based, in which context Instagram has a leading role. This trend is partly attributed to the proliferation of image creation and sharing on smartphones (Ibrahim, 2015, p.43; Stuart et al., 2017). As such, more and more libraries are employing Instagram to build up their profiles, market events, and communicate with library patrons (Hopkins, Hare, Donaghey, & Abbott, 2015, p.16).

Last but not least, Tufekci (2014) identifies a methodological issue of social media studies: the prevalence of single-platform studies that neglect the broader social influence in interaction and diffusion. Twitter is considered to be a disproportionately studied social media. Compared with Twitter, Instagram has received much less attention from academia. However, photo-oriented social media can disclose more feelings of individuals (Pittman & Reich, 2016), which demonstrates that Instagram is prior to Twitter to discover user patterns.

### 1.2.2 Contribution

The accomplishment of this doctoral study will make concrete contributions. From the theoretical point of view, there are research gaps between Big Data and public libraries, especially in Finnish public libraries. This doctoral study, focusing on helping public libraries realize what their responsibilities might be in the context of Big Data and to understand what Big Data is and how it can be applied in public libraries, will fill such gaps. In addition, the methodological issue identified by Tufekci (2014), concerning the prevalence of Twitter in social media studies, is avoided. Study 3 and Study 4 enrich the research content by analyzing Big Data from Instagram.

From the practical point of view, this study is motivated by the curiosity of librarians in Finnish public libraries. They do not know what Big Data is and how to apply Big Data in their daily work. As such, they would like to see a real-world example to help them understand Big Data and Big Data applications. The presented study can satisfy their requirements. Moreover, the fulfillment of Study 3 and Study 4 outlines effective communication ways with library users and realizes what current readers like or dislike. These outcomes can help (Finnish) public libraries boost the communication of active library users and provide more user-oriented services.

All in all, the presented study is significant for (Finnish) public libraries to understand and utilize Big Data.

## 1.3 Thesis structure

This dissertation summary is divided into six sections. The first section introduces the purpose, research questions, motivations and contributions, so as to provide basic knowledge of the research. The second section is the literature review section. The next section introduces the methodology to answer all the research questions, after which, the results of the studies are presented. The fifth section contains a discussion of the research findings, and the sixth and final section presents the conclusions and research limitations of the doctoral study.

# 2. Literature review

There are three major concepts in this doctoral study: Big Data, the public library, and Instagram. The public library is a component of the public sector, just as Instagram is a component of social media. The Library 2.0 paradigm brings social media to public libraries, with various benefits as listed in Table 1. A great amount of data has been generated on social media such as Instagram. Wisely applied, Big Data could be valuable for public sectors, e.g., the public library. Therefore, this section begins with a brief examination of the public library so as to outline the context of the study. The focus next turns to the definition of Big Data as studied or used in previous studies, which is reviewed in order to shed light on what Big Data is and to help to develop a definition of Big Data in librarianship in the present doctoral study. Studies on Big Data in the public sector and the public library are reviewed. Challenges and benefits of Big Data in these contexts are outlined, which are useful to identify roles of public libraries in the era of Big Data. Since Instagram is the chosen source for collecting Big Data for this study, the applications of social media in the public library are first summarized and the use of Instagram in librarianship is then introduced. Since this study chooses hashtags as the angle from which to explore Big Data applications for public libraries, the meaning, benefits, and use of hashtags are reviewed as well.

## 2.1 The origin and meaning of the public library

Libraries are the representation of essential, cherished, symbolic, and cultural resources (Harris, 1999, pp.3-7). The development of libraries has been supported by religious, political, and social ideologies so as to reach desirable religious, political, and social ends. For example, the emergence of the concept of democracy advocates the participation of civilized citizens during the process of political decision-making, but the spread of democracy requires free and equal access to information for all citizens, which in turns lays the foundation for the construction of public libraries (Harris, 1999, p.6).

Harris (1999, pp.149-161) states that public libraries emerged in the eighteenth and nineteenth centuries, which could be attributed to the establishment of a new society based on access to knowledge. Harris (1999, p.149) conceptualizes the public library as the "general library that is not only publicly owned and tax-supported, but also open to any citizen who desires to use it." Mandel (2013, p.264) describes a public library as a physical manifestation owning various materials, devices, and furniture to provide services for the community freely. With the Internet spreading worldwide, public libraries start to meet various information needs of citizens, and providing access to the net has become the key element of the service in the public library (Jaeger, Greene, Bertot, Perkins, & Wahl, 2012). Public libraries are also considered networked organizations that connect

their staff, collections, and physical spaces to citizens and various communities (Cavanagh, 2015; Hicks, Cavanagh, & VanScoy, 2020), which demonstrates the feasibility of social media applications for public libraries.

The development of information technology has changed peoples' ways of life. Public libraries are in a position to involve such cutting-edge technologies and educate people about the use of these technologies (Hoy, 2014; Jaeger et al., 2012; Ylipulli & Luusua, 2019), which implies that public libraries have a responsibility to understand the meaning and application of Big Data as a rising concept and bring it to people's lives. However, there are no studies identifying the specific roles that public libraries should undertake in the context of Big Data.

Finland has the highest number of active library users across the European Union (Prior, Toombs, Taylor, & Currenti, 2013). Based on a study by Vakkari et al. (2014), compared with Norway and Netherlands, both of which are small or medium-sized states and have similar welfare as Finland, Finnish public libraries are open longer, allowed more operational costs, and staffed with more professional librarians. Finnish public libraries have been considered successful democratic projects and enjoy a positive reputation domestically and globally (Ylipulli & Luusua, 2019). Therefore, this study is organized in a situation where public libraries are highly developed.

## 2.2 The definition of Big Data

Big Data was first mentioned in a study by Cox & Ellsworth (1997), who explained the problem caused by Big Data for visualization algorithms. According to studies by De Mauro, Greco, & Grimaldi (2015) and Li, Tao, Cheng, & Zhao (2015), the advent of Big Data is the outcome of an information explosion. This is vividly explained by Kumar & Sing (2015), who point out that in the year 2006 alone, the amount of information generated is equal to three times the information in all books ever written. At the beginning of the 21st century, Lyman & Varian (2000) realized that the world was heading towards a sea of information. They argue that it is wise to take full advantage of such a supply of information with better tools and better comprehension. Information technologies, such as information collecting, information storing, and information processing are highly developed, which boost the wave of Big Data as well (Bryant, Katz, & Lazowska, 2008; Li et al., 2015; Moise, 2016; Xu, Cai, & Liang, 2015). The explosion of information and the development of information technologies bring new opportunities and requirements for individuals and organizations. From a business perspective, Laney (2001) highlights that novel data management approaches are necessary owing to the surge of e-commerce, the rise in database merge and collaboration, and the drive for harnessing information. Laney puts forward three dimensions of data management:

- Volume: the depth or breadth of data.

- Velocity: the pace at which data is generated or used to support transactions.
- Variety: various formats, structures, and semantics of data.

These three V's lay the foundation for future studies to define Big Data. Bryant et al. (2008) outline that our world is heading to a data-driven future and consider computing Big Data as one of the biggest innovations in the past decade. From then on, Big Data has been widely discussed in academia and industries.

A consensus on the importance of Big Data has been reached, but the definition of Big Data has not been agreed upon widely. Instead of summarizing the meaning of Big Data with sentences, scholars tend to use themes or traits to define Big Data. Katal, Wazid, & Goudar (2013) define Big Data with six traits: variety, volume, velocity, variability, complexity, and value. The first three traits share the same meaning as the three dimensions of data management put forward by Laney (2001). Variability refers to the inconsistencies of data loads. For instance, more posts would be made on social media with special events coming soon. Complexity signals that it is difficult to clean, maintain, and transform the various types of data from diverse origins. Value represents the insights or patterns attained via analyzing Big Data. Ultimately, Katal, Wazid, & Goudar (2013, p.404) conceptualize Big Data as "large amount of data which requires new technologies and architectures so that it becomes possible to extract value from it by capturing and analysis process." De Mauro et al. (2015) employ four themes to comprehend Big Data: (a) information, (b) technology, (c) methods, and (d) impact. They propose a definition of Big Data as "information assets characterized by such a High Volume, Velocity, Variety to require specific Technology and Analytic Methods for its transformation into Value (p.103)."

Inspired by Laney, many scholars use V's to define Big Data. In order to provide a broader understanding of Big Data and embrace its essence, Al Nuaimi, Al Neyadi, Mohamed, & Al-Jaroodi (2015); Erevelles, Fukawa, & Swayne (2016): Gandomi & Haider (2015); and Gartner (2013) all use three V's (volume, velocity, variety) to explain Big Data. Gandomi & Haider (2015) and Erevelles et al. (2016) agree that additional dimensions could be added to refine the definition of Big Data. When demonstrating the benefits and challenges of Big Data, Buhl, Röglinger, Moser, & Heidemann (2013) use four V's to represent Big Data: volume, velocity, variety, and veracity. Veracity is relevant to one of the six traits of Big Data put forward by Katal, Wazid & Goudar (2013): complexity which implies the challenge of managing data when the amount reaches the peak. This four-V model has been widely used in various studies (Abbasi, Sarker, & Chiang, 2016; Dong & Srivastava, 2013; L'heureux, Grolinger, Elyamany, & Capretz, 2017).

With the development of new studies in Big Data, another V has been added to explain Big Data: value (Rajaraman, 2016; Storey & Song, 2017; Yin

& Kaynak, 2015). Value represents the benefits of analyzing Big Data to improve services, refine strategies, understand users, and so on. According to the study by Storey & Song (2017), value is the center of the other four V's. By reviewing previous studies, Saggi & Jain (2018, p.764) come up with an additional two V's: valence (data connectivity) and variability (the constant change of data meaning).

Apart from these V's, other attributes are also used to define Big Data. According to the review of Kitchin & McArdle (2016), exhaustivity (including the overall dataset rather than a sample of it), fine-grained in resolution and uniquely indexical in identification, relationality (the capability of connecting other datasets), extensionality and scalability (the data can be modified and extended) are all useful features of comprehending Big Data. Owais & Hussein (2016) develop nine V's as Big Data characteristics, which are included in five processes:

- Data collection: variety and veracity;
- Data processing: velocity and volume;
- Data integrity: validity, variability, and volatility;
- Data visualization: visualization
- Data value: value

The V's of data integrity and visualization are added to the five-V model. Data integrity is regarding the precision and consistency of Big Data. Visualization refers to the necessity of understanding Big Data in a graphic way. These nine V's of Big Data not only provide a thorough way to comprehend Big Data but also guide organizations to use Big Data in a systematic manner.

The aforementioned studies define Big Data from different perspectives, which indicates that Big Data is more than a buzzword. However, there are few studies concentrating on comprehending Big Data in the context of a specific field or industry, which enriches the significance of the present doctoral study in putting forward a definition of Big Data within the context of the library. Such a definition considers the characteristics of Big Data generated in librarianship and could be helpful for librarians to further understand and utilize Big Data.

## 2.3 Big Data in the public sector

Although a consensus has not have been reached on the definition of Big Data, studies have been conducted to explore the potentials of applying Big Data in various fields. The value creation aspect of Big Data has been realized in a business context after companies witnessed the increasing generation of data (Banica & Hagiu, 2015; Wamba et al., 2015). According to the review by Wamba et al. (2015), the application of Big Data in a business environment can be grouped into four major areas: (a) replacing/supporting human decision-making with automated algorithms;

(b) discovering needs, exposing variability, and improving performance; (c) innovating business models, products, and services; and(d) customizing actions.

Through the application of Big Data, companies can innovate products, better understand customers, and better handle employee-related issues (Banica & Hagiu, 2015). These benefits demonstrate that Big Data is a valuable resource for companies and could encourage the development of more varied ways to utilize Big Data in the near future. The application of Big Data is not only limited to a business environment: it is also widely employed in the public sector. Klievink, Romijn, Cunningham, & de Bruijn (2017) classify Big Data applications in the public sector into three types: object evaluation, research (seeking new insights), and continuous monitoring. Big Data applications show substantial potentials in saving resources, providing sustainable services, and attaining high efficiencies (Desouza & Jacob, 2017; Kim, Trimi, & Chung, 2014; Malomo & Sena, 2017).

In this section, Big Data applications in the public sector are first reviewed, and then Big Data applications in the public library. This section thus provides a deeper understanding of how Big Data has already been applied in the public sector and in the public library.

### 2.3.1 Big Data applications in the public sector

One of the characteristics of Big Data has been widely accepted: value, according to the reviews by Fredriksson et al. (2017), Gul & Ahsan (2019), and Okwechime, Duncan, & Edgar (2018). With the increasing availability of data and the development of methods to use data, the public sector has become more and more aware of the importance to seek for the benefits of Big Data (Maciejewski, 2017, p.121). Exploratory studies have been conducted on Big Data benefits for the public sector. These benefits can be broadly classified into six themes, which represent the main application fields for Big Data in the public sector:

1) Predicting social and economic status (Fredriksson et al., 2017; Gamage, 2016; Malomo & Sena, 2017; Patel, Roy, Bhattacharyya, & Kim, 2017). The greater the amount of data is, the more accurate the prediction could be.
2) Improving policy and decision making (Fredriksson et al., 2017; Gamage, 2016; Maciejewski, 2017; Malomo & Sena, 2017; Okwechime et al., 2018; Patel et al., 2017; Vydra & Klievink, 2019). Big Data helps realize the transformation of policy and decision-making from intuition- or experience-based to data-driven, which innovates the process of identifying problems
3) Increasing operational efficiency (Patel et al., 2017; Sarker, Wu, & Hossin, 2018). Key factors of a project or a process can be identified via Big Data analytics, thus increasing the operational efficiency. For instance, by analyzing data

concerning vehicles, roads, and citizens mobility, public transportation can be arranged wisely during rush hour.

4) Increasing transparency (Gamage, 2016; Sarker et al., 2018). Through visualizing great amounts of data, the transparency of a public event could be increased so as to boost the level of understanding between citizens and public sectors. For instance, during the ongoing COVID-19 pandemic, the public health sector publishes various figures to illustrate the spread of this infectious disease.
5) Improving services (Gamage, 2016; Maciejewski, 2017; Malomo & Sena, 2017; Patel et al., 2017; Sarker et al., 2018; Vydra & Klievink, 2019). Patterns and behaviors of citizens can be outlined by analyzing Big Data; these are significant for public sectors to deliver more customized services according to the patterns of different people.
6) Lowering costs (Gamage, 2016; Sarker et al., 2018; Yigitcanlar, Desouza, Butler, & Roozkhosh, 2020). With the application of artificial intelligence (AI) techniques on big datasets, public sectors can identify risks, save resources, and facilitate working processes, which decreases costs significantly.

In order to ensure these benefits of Big Data, studies have been carried out to put forward suggestions on the efficiency and effectiveness of Big Data applications for public sectors. Maciejewski (2017) defines three administrative functions of Big Data applications in public sectors by reviewing previous studies: (a) public supervision for identifying irregularities, (b) public regulation for situational awareness and feedback and, (c) public service delivery. Maciejewski also discusses the requirement to guarantee these administrative functions of Big Data application. For public supervision, since irregularities are planned to be detected, therefore enough data for analysis, models to identify irregularities and technologies and device to apply the model on the data are needed. For public regulation, the techniques to gather real-time data and present the result of data analysis are significant for giving feedback. As for public service delivery, it is important to interpret the data from a customer behavior point of view and to understand the essence of the service.

By interviewing twelve experts in information systems and technologies, Merhi & Bregu (2020) rank thirteen factors leading to the effective use of Big Data in public sectors. These factors can be classified into three groups: techno-centric factors (seven factors), user-centric factors (three factors), and moderate factors (three factors). Techno-centric factors rank the highest compared with the other two classes. Among techno-centric factors, the factor of authentication is the most important, which ensures the use of government data among different parties. All factors concerning security control of techno-centric factors are at the top of the ranking list, which

further demonstrates that confidentiality, privacy, and security control are foremost for public sectors in using Big Data (Gamage, 2016). Collaboration between various public sectors or between public and private sectors is encouraged so as to strengthen the infrastructure and network, to share experience, and to fasten the process of the application (Gamage, 2016; Malomo & Sena, 2017; Okwechime et al., 2018; Sarker et al., 2018). To ensure the success of Big Data applications in public sectors, it is also useful to help personnel acquire necessary skills (Gamage, 2016; Malomo & Sena, 2017) and to gain support from senior management (Malomo & Sena, 2017). Although Big Data is a valuable resource for organizations, there are challenges that come with its application. When it comes to Big Data, public sectors need to confront three core dilemmas: (a) whether to change current practices as little as possible during the implementation of Big Data, which is echoed by the study of Klievink et al. (2017,p.272); (b) whether to take the risks and uncertainties of involving Big Data; and (c) whether to make the transformation technology-based or human-based while developing the application of Big Data. These dilemmas hinder the public sector from embracing Big Data-related transformations (Kuoppakangas et al., 2019).

The attitude of some managers in public sectors complicates the situation as well. Guenduez, Mettler, & Schedler (2020) interviewed 32 managers working in public sectors on their opinions towards Big Data. Some of them were skeptical about Big Data, did not trust Big Data, or showed little enthusiasm concerning Big Data. Some held neutral opinions towards Big Data, but they observed no benefit of utilizing Big Data either. The existence of such public managers makes the involvement of Big Data in public sectors challenging. In addition, the abilities of the sector in collecting, archiving, and retrieving data; processing data; visualizing data; interpreting the analysis results; protecting privacy; encouraging data sharing and providing proper access; and training personnel to obtain key skills are also challenges (Bram et al., 2017; Fredriksson et al., 2017; Sarker et al., 2018).

### 2.3.2 Big Data applications in the public library

Even though Big Data is not new in the public sector, the application of Big Data is not evenly distributed in each public sector (Fredriksson et al., 2017). Few studies have been conducted in the context of the public library, which is the key sector to support the spread and use of technologies developed since the advent of Big Data (Ylipulli & Luusua, 2019). According to the DIKW (data, information, knowledge, and wisdom) hierarchy, data is transformed into information, information is transformed into knowledge, and knowledge is finally transformed into wisdom (Rowley, 2007). Public libraries, as hubs of information, are responsible for providing knowledge to users so as to support lifelong learning, peace, cultural development, and social welfare (Abumandour, 2020). That is to say, public libraries should organize data, information, and knowledge to meet such responsibilities.

Therefore, it is inevitable for public libraries to manage, and obtain information and knowledge from Big Data.

Nowadays, the research concerning Big Data in the public library mainly focuses on how to employ Big Data for better strategies, service, and policy-making in public libraries. Zou, Chen, & Dey (2015) analyze more than 10,000 tweets for ten big public libraries in US to highlight how users engage with the public library. They classify four engagement method categories based on the purpose of engagement: literature exhibits, engaging topics, community building, and library showcasing. They notice that certain libraries might emphasize one of the engagement methods in their daily tweets. Tweets regarding literature exhibits and engaging topics attain more user feedback compared to tweets regarding the other two engagement methods. Such findings help public libraries come up with good strategies to engage with their users via Twitter. Kim & Cooke (2017) employ the Chernoff face method to analyze statistical data on library operations, services, and users to the end of 2014 in order to compare the performance of public libraries in London and Seoul. Chernoff face method uses the size of facial features to represent the key figures regarding library operations, services, and users. For instance, the wider the hair is, the more visits the public library receives; the bigger the eyes are, the more collections the public library owns; and so on. The uniqueness of each face represents the operations, services, and visiting situation in each library. In the end, they created 36 faces of public libraries in London and 28 faces of public libraries in Seoul. After comparing these faces, they conclude that public libraries in London receive more support from local government than those in Seoul, leading to better public library operations and services in London than in Seoul. The result also reflects that Seoul needs more public libraries, collections, and budgets to enhance library use. Crawford & Syme (2018) present cases of analyzing hundreds of thousands circulation data to enhance the collection management in the public library with the help of collectionHQ, which is a Big Data analytic tool supporting library management. The use of collectionHQ helps librarians to prepare materials meeting user demand, maximizes the usage of current collections, promotes collections, and monitors the circulation of collections. Chang (2018) has developed a platform to trace the history of the Hakka people living in Taiwan. He combined Big Data technology and GIS mapping technology to facilitate the family tree visualization of Hakka people by analyzing 4492 Hakka genealogies covering over 2000 years. Such a platform indicates that Big Data technologies generate suitable tools to improve library service. Kim, Gang, & Oh (2019) analysed 96,086 pieces of user data and 3,361,284 logs to identify the factors of the spatial usage in the public library so as to provide better operation polices.

Based on the review, public libraries have noticed the benefits of applying Big Data in their daily services since the year 2015. However, there are not

many public libraries utilizing Big Data in practice to date, which is reflected by the limited number of reviewed studies in this field. Furthermore, it can be concluded that current studies have been focusing on using Big Data for operations evaluation, service improvement, collection management, and decision-making. Fewer studies focus on employing user-generated data on social media to understand users' reading needs or to enhance communication with users, which indicates a lack of research. There are also fewer studies that put Big Data in the context of public libraries and demonstrate the roles and responsibilities of public libraries to meet the advent of Big Data. This doctoral study will fill these research gaps and enlarge the domain of Big Data applications in the context of public libraries. In addition, the data in the aforementioned studies concerning Big Data applications for public libraries is either much less than one million or historical. Therefore, the present study decides to use greater amount of data generated on social media, which is less historical and more real-time, to enrich the options of data that can be applied for public libraries.

## 2.4 Social media in the public library

The utilization of social media in the library context can be referred to as the movement of Library 2.0, "a change in interaction between users and libraries in a new culture of participation catalyzed by social web technologies"(Holmberg, Huvila, Kronqvist-Berg, & Widén, 2009). Library 2.0 requires the abilities of libraries to function in the context of social media (Huvila et al., 2013, p.204). Kronqvist-Berg (2014) states that public libraries are positive about user interests in social media and the inherent benefits of user participation through social media. Studies have been conducted to seek the benefits of using social media in public libraries. The practices of social media in the public library are summarized in Table 1:

**Table 1: The summarization of social media practices in the public library**

| | Practices |
|---|---|
| **For library management** | **Staff development** (Hall, 2011) |
| | **Internal communication** (Vanwynsberghe, Boudry, Vanderlinde, & Verdegem, 2014) |
| **For library service** | **Marketing** (Abidin, Kiran, & Abrizah, 2013; Fernandez, 2009; Hall, 2011; Huang, Chu, & Chen, 2015; Islam & Habiba, 2015; Kaushik, 2016; Lamont & Nielsen, 2015; Mon & Lee, 2015; Vanwynsberghe, Vanderlinde, Georges, & Verdegem, 2015; Vassilakaki & Garoufallou, 2015, 2014; Xie & Stevenson, 2014) |
| | **Information sharing** (Abdullah et al., 2015; Abidin et al., 2013; Cahill, 2011; Fasola, 2015; Hall, 2011; Phillips, 2011; Xie & Stevenson, 2014; Young, 2016) |
| | **User communication** (Abidin et al., 2013; Carlsson, 2012; Fasola, 2015; Fernandez, 2009; Ganster & Schumacher, 2009) |
| | **Relationship establishment** (Abidin et al., 2013; Cahill, 2011; Carlsson, 2012; Young, 2016; Young & Rossmann, 2015) |
| | **Providing customer services** (Abdullah et al., 2015; Anwyll & Chawner, 2013; Cahill, 2011) |
| | **User feedback collection** (Cahill, 2011; Xie & Stevenson, 2014) |
| | **Service renewal** (Fasola, 2015; Gan, 2016; Hall, 2011; Vassilakaki & Garoufallou, 2015) |

The category of social media practices can be classified into two main groups: for library management and for library services. From a library management perspective, social media are mainly helpful in staff development and internal communication. According to Hall (2011, p.422), various videos of training courses can be found on YouTube, which are beneficial for librarians to improve their professional skills. While accomplishing a program aiming at increasing the social media knowledge and skills of librarians, Vanwynsberghe et al. (2014) noticed that internal communication has been facilitated by social media.

As seen in Table 1, most social media practices in the public library are associated with library services. Among these practices, marketing has been

done very often through social media. This is reflected by the survey conducted by Islam & Habiba (2015) on the use of social media in Indian libraries. In the end, 82.5% responding librarians claimed that they use social media to market library services and products. New services, new resources, upcoming events, exhibitions and activities, and digital archives are marketed through social media as well.

Social media are also used for information sharing. Xie & Stevenson (2014) identified that information sharing has been conducted on social media of the studied digital libraries in the manner of sharing links, digital or physical collections. In describing the utilization of Web 2.0 tools at Vancouver Public Library (VPL), Cahill (2011) provided an example of information sharing through social media. When there was a power breakdown in downtown Vancouver, VPL added a post on this matter on their Twitter homepage. Such a pattern has been proved to be an effective information sharing method when there is an emergency. Cahill summarizes that VPL also employs Twitter and YouTube to generate excitement about new branches to get engagement with communities. Such relationship establishment is also reflected by the study of Carlsson (2012) and Abidin et al. (2013). Focusing on the interplay between Facebook and the librarianship in a Swedish public library setting, Carlsson (2012, pp.207-208) observed that Facebook homepages of the public library are designed to build relationships with library users. Abidin et al. (2013, pp.83-84) examine the adoption of Web 2.0 applications in Malaysian public libraries through face-to-face interviews and an online survey. The result of the study shows that one main use of social media in Malaysian public libraries is social networking.

Collecting user feedback is also a common practice, which is exemplified by Fasola (2015, p.871) when discussing the perception and acceptance of librarians towards the use of Facebook and Twitter in promoting library services. This practice is reflected by Cahill (2011, p.266) as well. Intending to solicit user feedback about a new library catalogue, one librarian at VPL posted on Twitter to encourage feedback. Eventually, two dozen responses came in the first few days.

Social media practices have a positive influence on library service innovation and renewal. From the viewpoint of Hall (2011), reference services can be extended beyond physical desks with the help of social media, because social media provide users with an instance access to communicate with librarians. This viewpoint is shared by Fasola (2015) and Vassilakaki & Garoufallou (2015), who argue that social media offering new ways to provide library services. When exploring the current status of WeChat (one of the most popular social media apps in China) application in Chinese public libraries, Gan (2016, pp.632-634) summarizes that custom menus are created specifically for individuals through WeChat, which implies a better solution to serve library users.

Holmberg et al. (2009, p.677) define interactivity as the most important part of Library 2.0. Based on the review above, it can be concluded that social media facilitate the interaction between libraries and users in marketing, information sharing, user communication, relationship establishment, providing customer service, user feedback collection, and service renewal. Therefore, it can be concluded that social media play an important role in Library 2.0. Among all the social media used by public libraries, Twitter and Facebook are the two most popular, as identified by Abdullah et al., 2015; Hussain, 2015; and Xie & Stevenson, 2014.

These practices reflect the benefits of using social media. After reviewing the use of Web 2.0 services in Polish urban public libraries, Wojcik (2015, p.100) concludes that librarians have learned how to understand user preference through social media. Such knowledge could be helpful to provide users with more relevant information. Book recommendation on social media sites is one good example to show the potential of social media to know library users deeply (Anwyll & Chawner, 2013). Cavanagh (2016, p.251) investigates micro-blogging practices in Canadian public libraries through an online survey. In the end, the result shows that the top three benefits of using micro-blogs like Twitter are facilitating internal staff management, gaining library awareness, and interacting with users. The benefit of gaining library awareness is also highlighted by Fasola, 2015; Fernandez, 2009; Huang et al., 2015; Islam & Habiba, 2015; and Kaul, 2016. Interaction with library users has been considered one of the obvious benefits of social media in various studies (Abdullah et al., 2015; Fasola, 2015; Fernandez, 2009; Kaul, 2016; Kaushik, 2016; Lamont & Nielsen, 2015; Xie & Stevenson, 2014). Apart from these benefits, Kaul (2016) also puts forward other benefits of social media, which are keeping updated with the most current information and developments in the profession, spreading information in a short time, facilitating decision-making, and enabling cross-national study and mobility.

Even though social media bring about benefits, there are still concerns. Staff-related issues such as lack of motivation, lack of skills, and productivity decrease are the main obstacles to the development of social media in the public library (Abdullah et al., 2015; Islam & Habiba, 2015; Kaul, 2016; Kaushik, 2016; Zyl, 2009). Since everyone can generate content on social media, authentication of information should also be a concern (Islam & Habiba, 2015). Furthermore, the easy access to social media inevitably leads to more malware created by spammers and virus-writers (Zyl, 2009). The majority of malware causes data leakage which brings about challenges to protect user privacy (Fernandez, 2009; Huang et al., 2015). Solutions to face such challenges should be considered before public libraries make any movement about social media. In addition, the lack of guidance and policies to macroscopically instruct public libraries to employ social media could also be an issue (Cavanagh, 2016; Vassilakaki & Garoufallou, 2015).

On the basis of this aforementioned review, it can be concluded that the development of technology, the changing lifestyle of individuals, and the responsibility of libraries jointly contribute to the utilization of social media in the public library. Although certain issues concerning social media need attention, the various practices and benefits of social media indicate that the application of social media in the public library could be expanded in the future.

## 2.5 The use of Instagram in librarianship

Instagram is an image-based application designed for smartphones. Via Instagram, users can make social activities by generating visual and textual contents (Zappavigna, 2016, p.272). With the help of Instagram, users can capture and share their life conveniently (Hu, Manikonda, & Kambhampati, 2014, p.595). Instagram was first released on October 6th, 2010, bringing a new approach to mobile photography. The launch of Instagram corresponds to the desktop photograph application: Flickr (Manovich, 2016, p.11). Nowadays, Instagram has been growing to be one of the most popular social media in the world. According to the report by West (2019), there are 500 million users accessing to Instagram every day, who spend 53 minutes daily on average. Instagram has 1 billion active users monthly. There are 100 million photos uploaded on Instagram every day. It can be concluded that Instagram has achieved tremendous success in its first decade. Millions of photos generated per day means a great amount of data created on Instagram, which provides a scenario of Big Data generation.

According to Manovich (2016, pp.11-12), Instagram is the most purely visual platform, merging various elements of photography (e.g., camera, photo paper, darkroom, exhibition, and publication). The increasing popularity of Instagram is not only attributed to the user-friendly features but also to the proliferation of image-intensive social software. According to the study by McNely (2012), Instagram is at the forefront of the current image-intensive social media. Apart from image-related features, Instagram includes significant features used on other social media as well, such as liking, commenting, and follower relationships. The interaction of Instagram users is both visual and textual. McNely advocates that organizations should consider photo sharing on Instagram more strategically, since professional management of activities on Instagram could be helpful to shape the image of an organization.

The use of Instagram in librarianship can be classified into four aspects: user engagement, library showcasing, keeping content dynamic, and possibility for fundraising.

### 2.5.1 User engagement

Libraries use Instagram to engage with users in three ways: liking user posts, reposting, and generating interactive content. Based on the working

experiences of Salomon (2013) in Powell Library, Instagram has a younger and more diverse user group than any other social media, which can be attributed to the fatigue of using other social media and the higher frequency to see posts that users are interested on Instagram compared with Facebook. Therefore, Powell Library was motivated to use Instagram for their library functions. They started by liking photos posted by students on Instagram that were taken in the library. With this approach, the number of followers starts to rise. The success of Instagram involvement in Powell Library inspired librarians in Indiana University's Herman B. Wells Library to create a geotag (#wellslibrary) that can help students automatically fill in the location where the photo was taken (Hild, 2014). In this way, user-generated content in Wells Library could be filtered easily by librarians so as to like the content. University of Liverpool Library, where Instagram have been showing a sharp rise recently, has also used Instagram to keep contact with students by liking their content on Instagram, (Chatten & Roughley, 2016). The purpose of reposting is twofold: to interact with the library users and to present the library diversely. Powell Library and University of Liverpool Library reposted certain photos posted by students, which gradually encouraged students to tag the library while posting. The more the library is tagged, the more visible the library's Instagram account becomes. In the case of Wells Library, reposting photos from the students' Instagram accounts could help showcase different angles of the library not seen when posting images taken only by library staff, which not only enhanced user engagement but also indirectly boosted the library outreach. Powell Library generated interactive content to improve students' engagement in teaching and learning. A content of a curriculum was created to boost the circulation of a course. Images regarding a certain field of knowledge, such as a map of brain, were posted to ask students who could explain the meaning. When answering the question, follow-up knowledge about the brain and human body was posted. Through this, Instagram at Powell library could be the platform to communicate with students and also to share knowledge. Wells Library (Hild, 2014) set up a screen in the lobby to display outstanding user photos of the library on Instagram, increasing the awareness of the library account and encouraging students to engage with the account. Hild (2014) considers that a visual format could generate more user loyalty and engagement because the application of Instagram in Wells Library achieves positive results. Chatten & Roughley (2016), Hild (2014), and Salomon (2013) hold positive opinions on the use of Instagram for user engagement. They believe that with the development of the social media environment, more values of Instagram can be explored for librarianship.

### 2.5.2 Library showcasing

The employment of Instagram in Rice Library and Roesch Library demonstrates the power of Instagram in presenting libraries (Tekulve &

Kelly, 2013). In Rice Library, the neglected library buildings were opted for showcasing the library, since there were not many events going on in the library. Roesch Library used Instagram to present construction progress on the library buildings. Both libraries created fun content on Instagram, attracting the attention of students and also boosting engagement with them. University of Liverpool Library (Chatten & Roughley, 2016) also uses Instagram to share the environment and atmosphere of the library. As a case study, Pustakalana's Library in Bandung, West Java Indonesia, uses Instagram to promote the library (Azwar & Sulthonah, 2018). Based on this study, Facebook and Instagram are the most popular social media among Internet users in Indonesia. Moreover, Azwar and Sulthonah insist that an increasing number of people tend to comment and judge via visual platform. Therefore, Instagram was selected to promote Pustakalana's Library. Pustakalana was considered as a library brand, and hence all the collections and events in the library were recognized as products of the brand. Information concerning the library products is shared via Instagram. Popular books, new collections, workshops in the library, quotes from a certain popular book, the guide on how to register the library website, and the opening hour of the library are all posted on Instagram. Librarians try to respond to the comments of each post to maintain an active communication with the followers as well. Consequently, the number of monthly visitors is increased which can be acknowledged as the fact that the awareness of the library and its services are promoted. The connection between the library and the patrons is strengthened, which is shared by the above studies in this section. However, it is also important for libraries to monitor the quality of the content on social media. Harrison, Burress, Velasquez, & Schreiner (2017), who reviewed the messages posted on social media of six different public and private university libraries, recommended that academic libraries should appoint librarians to monitor the social media account, not only to ensure the quality of the content on a regular basis but also to guarantee the spelling, grammar, and punctuation of the message, since a post on the social media represented both the library and the university. Therefore, when using Instagram for showcasing the library, the content of the post should be carefully checked both textually and visually.

### 2.5.3 Keeping content dynamic

The Mansfield Library Archives and Special Collections (ASC) employs Instagram to keep active content on their homepage (Wilkinson, 2018). The Instagram account of ASC was launched in 2015, which was primarily for communicating with fellow libraries. With vivid images posted by the account, ASC is followed by thousands of individuals and organizations. This motivated the librarians of ASC to embed Instagram on their homepage to get dynamic content. Based on Wilkinson's observation, the more dynamic the web content is, the more delighted users could be, because dynamic

content is established by visually appealing and interesting information. Such dynamic content can also be valuable for librarians, who could have a chance to look into the good collections storing in the library for a long time and come up with fresh ideas to circulate such cultural heritage. This process can be achieved by one person in a short time, which is significant for libraries currently confronting resource-limiting situations. To sum up, the practice as put into action at ASC demonstrates the affordability and applicability of Instagram for generating dynamic web content.

### 2.5.4 Possibility for Fundraising

According to Garczynski (2017, p.112), "Instagram is the social media platform that best enables libraries to demonstrate the strength of their resources and services." The successful promotion of Pustakalana's Library proves this statement. Furthermore, Garczynski thinks that Instagram is useful for fundraising. Even though there is no actual case of a library using Instagram to raise funds, the former successful cases imply the suitability. Garczynski suggests three steps for fundraising via Instagram:

- Setting a fundraising goal and reporting the progress.
- Thanking the donor publicly on Instagram and helping boost awareness of the donor, especially when the donor represents an organization.
- Posting the positive aspects of the donation.

The geotag is recommended in fundraising posts on Instagram, because outlining the location of the library is crucial for establishing a community, encouraging a favorable reception from locals. All in all, Garczynski perceives the potential value of Instagram to raise funds with its visual storytelling features.

All of the studies demonstrate that Instagram is worthwhile to be applied in the library context. Its liking and reposting features make communication with library patrons convenient, thus enhancing engagement with them. Its visual contents make it possible to present library environments vividly so as to showcase libraries, while its dynamic contents help libraries realize their own collection situation and thereby allocate recourses effectively. However, the current studies focus less on the hashtags generated for visual content on Instagram, which could also be valuable resources for public libraries with proper analysis. Likewise, there are few studies on Instagram applications for libraries from a Big Data point of view. Furthermore, most studies involve the application of Instagram in academic libraries. When it comes to studies on the public libraries, the number is much smaller. Therefore, the present study applies quantitative analysis on hashtags in captions collected from Instagram and outlines insights from the analysis of such textual contents to improve public library services.

## 2.6 Hashtags on social media

Hashtags are a basic feature of Instagram (Sheldon & Bryant, 2016), but it is also significant that the number of hashtags on the Internet has been increasing (Lee, 2016). Hashtags first appeared in conversations between the Internet Relay Chat application and its users (Giannoulakis & Tsapatsoulis, 2015) and have since become widely used in social media. According to the studies on hashtags by Dumbrell & Steele (2015), Fink et al. (2016), and Gotti and Langlais, & Farzindar (2014), hashtags refer to words, phrases, and alphanumeric characters prefaced by the symbol '#'. Speaking broadly, the functions of hashtags are (a) to denote the semantic content of the text (Burgess, Galloway, & Sauter, 2015; Giannoulakis & Tsapatsoulis, 2015; Ma, Sun, & Cong, 2013) and (b) to communicate within a community (Burgess et al., 2015; Small, 2011). Building on these two functions, current studies have been focusing on the use of hashtags to understand concepts and domains and to boost communication.

Dwyer & Marsh (2014) collected and analysed tweets with #trust. They outline the most used words with trust, thus helping to design a trust interaction interface. Pinho-Costa et al. (2016) analysed tweets by calculating Hashtag Global Reach, Topic-Specific Hashtag Global Reach, and Individual Global Reach, using these to generate an index of hashtags relevant to primary care and family medicine. They ranked six related hashtags and identified themes with global visibility. Rich, Haddadi, Hospedales, & Acm (2016) identified the hashtags most used in food-related images so as to help researchers understand current diet habits. All of these studies have employed hashtags as the perspective to comprehend certain concepts.

Due to the high searchability of hashtags (Dwyer & Marsh, 2014; Stathopoulou, Borel, Christodoulides, & West, 2017) and their rhetorical functions to reveal personal feelings and experiences (Burgess et al., 2015; Daer, Hoffman, & Goodman, 2014; Hamed & Wu, 2014; Rich et al., 2016; Small, 2011), hashtags are used to understand and encourage human behaviors. Small (2011) and Zappavigna (2015) believe that, wisely used, hashtags could increase an account's followers. Siapera (2014) identifies active people by analyzing tweets containing #Palestine. Chae (2015) summarizes that people who use #supplychain in their tweets are more conventional and information-focused. Gibbs, Meese, Arnold, Nansen, & Carter (2015) find that people tend to reveal their affective context and reposition their funeral experiences by analyzing images tagged with #funeral collected from Instagram.

When it comes to the communication function of hashtags, Cooper (2016) advocates that each tweet should use one or two hashtags but argues that it is not more useful to have more than two hashtags in a tweet to attain more likes and comments. Lee (2016) submits that the best number of hashtags in

an Instagram caption is eleven from an interaction perspective. Oh et al. (2016) suggest that users should be allowed to add new meanings to a hashtag, which is called retagging. Retagging will enlarge the audience, thus improving the communication function of hashtags. Both Bunskoek (2014) and Harkai (2018) advise that the more relevance that a hashtag's meaning has for the visual content on Instagram, the more possibilities that the marketing goal could be achieved on Instagram. Bunskoek (2014) also recommends controlling the length of hashtags and creating unique hashtags for marketing activities on Instagram.

Although achievements have been made in research on hashtags analysis, most studies concern hashtags in the domains of public administration, risk management, communication, and business management. No study centers on analyzing hashtags in the context of public libraries. However, libraries are in an ideal position to understand hashtags, owing to their deep involvement in social media and their responsibilities in annotating electronic resources (Giannoulakis & Tsapatsoulis, 2015).

Another observation is that hashtag research has disproportionately focused on Twitter. Owing to the fact that Twitter is more open to data collection (Chae, 2015), most studies analyze only hashtags attained from Twitter. This, in combination with the uneven spread of research as described above, has informed the decision in the present doctoral study to provide practical examples of analyzing a large number of hashtags from Instagram for the benefit of public libraries.

# 3.Methodology

The following section addresses the research approach of this doctoral study. The motivation of applying the methods for the studies included in this dissertation (Study 1-4) is briefly explained. It also provides a presentation of the research design for each study in detail.

## 3.1 Research approach

There are two types of reasoning for scientific inquiry: deduction and induction (Dastrup, 2019). In deductive research, the statements in the "if P, then Q" format are provided. Conclusions will be drawn based on theories from previous studies (Krawczyk, 2018, p.201). The current research in this particular area, Big Data in the library context, cannot provide enough theories to support the presented doctoral study to do deductive research. As is discussed in Section 1.2.1, there are few studies focusing on the understanding and application of Big Data for public libraries. The disproportionate coverage of studies on social media indicates the lack of analyzing Big Data from Instagram, not to mention in the context of public libraries.

As for inductive research, knowledge is attained and generalized from examples. Induction helps research move forward in terms of new knowledge (Krawczyk, 2018), which indicates that induction suits research in its infancy. Therefore, the four studies included in this doctoral study, are developed from inductive approaches.

Scientific research can be classified as qualitative and quantitative. Qualitative research concerns the aspect of reality that cannot be quantified. The object of qualitative research is to generate in-depth information of various dimensions thus understanding the research question. Quantitative research draws conclusions via analyzing quantifiable variables and inferences. These two paradigms of scientific research are widely accepted (Queirós, Faria, & Almeida, 2017). The qualitative vs quantitative contest has been ongoing for years. Each research paradigm has its own strengths and weaknesses. It is recommended to consider the research question while deciding whether research is qualitative or quantitative (Onwuegbuzie & Leech, 2005; Queirós et al., 2017). Moreover, Onwuegbuzie and Leechn (2005) advocate the combination of these two methodologies, especially when the research data is limited.

This doctoral study is conducted by inductive approaches, combining qualitative and quantitative methodologies. The overall purpose of this study is to help public libraries realize what their responsibilities might be in the context of Big Data and to understand what Big Data is and how it can be applied. In order to fulfill this purpose, four studies have been conducted. The theoretical connection of these four studies is illustrated in Figure 5.

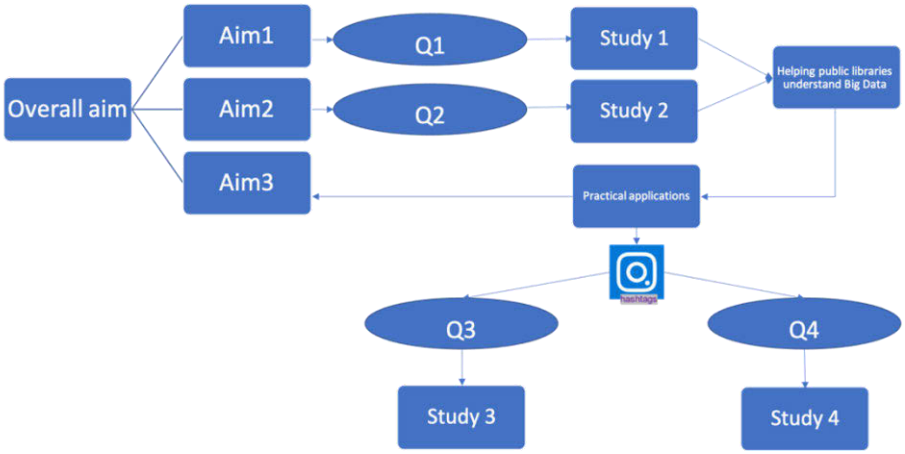



**Figure 5. The theoretical connection of four studies**

The overall aim is split into three subordinate aims. Therefore, the achievement of the overall aim can be accomplished by these three aims which are reached by answering four research questions in four studies as explained in Section 1.2. Moreover, the accomplishment of the first two aims helps public libraries understand Big Data, which guides Big Data applications in public libraries. Practical applications are created based on data collected from Instagram. In this doctoral study, hashtags are chosen for analysis, based on two main functions of hashtags: to communicate with the community and to denote the semantic content of the text.

The motivation of applying the method for each study is based on the corresponding research question:

Study 1: what kind of roles should public libraries undertake in the context of Big Data?

To answer this question, a list of roles is needed. Since there is no research on the roles of the public library in the context of Big Data, relevant roles to Big Data were first generated based on discussion with academic librarians with knowledge of Big Data, the suggestions from previous studies, and the characteristics of data saved in public libraries. Both quantitative and qualitative methods were used. Firstly, the level of agreement on the proposed roles was evaluated through an online survey with numerical measures. Secondly, semi-structured interviews were conducted with library directors to deepen comprehension of these roles. In the end, roles are identified by this combined method. The strength of the online survey and semi-structured interviews are explained below in Section 3.2.

Study 2: what does Big Data mean specifically in librarianship?

A definition of Big Data specific to librarianship is the aim to answer this research question. Since no prior studies focus on defining Big Data in a specific field, this study uses the definitions of Big Data presented in

academic literature in the field of library and information science (LIS) as the reference when summarizing key aspects of these definitions. These key aspects are then examined together to define Big Data specifically in librarianship. The summarization of key aspects is achieved by analyzing definitions through (a) statistical description of the definitions, which presents the frequency of keywords and the similarity of each definition based on the words used in the definition, and (b) content analysis, which has been used in LIS studies since the 1990s and is good for attaining the structure and patterns of texts (Mahraj, 2012). That is to say, Study 2 also combines quantitative and qualitative methods to ensure the quality of research.

Study 3: how should libraries effectively organize hashtags to attain more "likes" and comments for library-related posts on Instagram?

Study 3 is designed to boost the communication function of hashtags. A quantitative approach was employed to answer the research question. As such, quantifiable variables are needed. The hashtag communication function is measured by how many "likes" and comments a caption attains. In order to focus on the communication function of hashtags, Study 3 excludes other factors that could increase the number of comments and "likes", such as the content of the post, the identity of the poster as a private individual or an organization, and the number of the followers an account already has. Instead, this analysis of the communication function of hashtags concentrates on hashtags organization, i.e., how and where hashtags are distributed in the caption. Different ways of organizing hashtags are identified as the hashtag location, i.e., the general location of all hashtags in a caption. The concept of hashtag location is developed based on the study by Gotti et al. (2014). In order to make hashtag location computable, the index of a hashtag in a caption is identified. This lays the foundation for further calculations. Regression models are created to establish the connection between different hashtag locations and the number of comments and "likes". The selection of regression model is made based on the prerequisites of certain regression algorithms, which is revealed by the statistical distribution of data. Details will be explained in Section 3.4.

Study 4: what topics do current readers like or dislike?

A list of topics is needed to answer this question. One of the functions of hashtags: denoting the content of the text was outlined to identify topics that current readers like and dislike. Sentiment analysis was employed, because it can "analyse people's opinions, sentiments, evaluations, appraisals, attitudes and emotions towards entities such as products, services, organisations, individuals, issues, events, topics and their attributes" (Liu, 2012, p.12). The process of sentiment analysis falls under the domain of quantitative methods. Through sentiment analysis, readers' attitudes and emotions concerning content could be outlined. Thereby, what readers like

and dislike could be revealed by these attitudes and emotions. In order to ensure the quality of sentiment analysis, both opinion polarities (negative, neutral, and positive) and emotion classifications (love, joy, sadness, fear, anger, and surprise) were identified in Study 4. In order to create the model for classifying opinion polarity and emotion, the supervised machine learning approach was applied. Machine learning concerns the process of modification and adaptation. Machine learning allows computers to adjust their performance via the induction or compilation of knowledge so that performance is improved (e.g., better prediction accuracy or better control of robot) (Marsland, 2014, pp.4-5; Sison & Shimura, 1998, p.134). There are four types of machine learning (Marsland, 2014, p.6):

- Supervised learning. Learning from examples. A training dataset with the correct responses (truth) is provided. Algorithms learn the pattern from this training dataset and forecast the response to all possible inputs.
- Unsupervised learning. Identifying similarities between inputs.
- Reinforcement learning. This is a dynamic learning process. The algorithm works interactively with the situation. The algorithm is informed of what is wrong, but it does not know what is right. It needs a process of learning to know how to get the correct answer.
- Evolutionary learning. This type is derived from biological evolution. The algorithm tries to figure out the best solution to fit into the situation where a score is given to denote how good the solution is.

In Study 4, sentiment analysis is conducted by supervised machine learning algorithms, because both opinion polarity model and emotion classification model are established based on a dataset with truth, which falls under the scope of supervised machine learning.

In summary, the application of methods for these four studies follows the suggestions by Onwuegbuzie and Leechn (2005). Study 1 and Study 2 have limited data, therefore qualitative methods and quantitative methods are combined to ensure the research quality. As for Study 3 and Study 4, the amount of data falls under the domain of Big Data, hence these two studies are approached by quantitative methods.

## 3.2 Research design of Study 1: Outlining roles of public libraries in the context of Big Data

Study 1 combines two methods: the online survey and the semi-structured interview. According to the study by Evans & Mathur (2005), there are sixteen strengths of online surveys. Twelve of them are relevant to the implementability of Study 1, as is presented in Table 2:

**Table 2: Strengths of online surveys and the meaning for the study**

| Strength | Meaning |
|---|---|
| Global reach | Since Big Data is a new topic in librarianship, the wider we can reach, the better insights could be attained for further identify the roles; |
| B to B or B to C appeal | In the study, no opinions are gained from the organization point of view, therefore this strength is not considered; |
| Flexibility | Owing to the way that the questionnaire can be accessed by clicking the link, ways to invite participants are flexible, which enlarges the possibility to get sufficient answers; |
| Effectiveness | Less time spent on receiving and managing responses; |
| Technological innovation | Questions will be presented on the screen rather than printed on paper, thus saving resources; |
| Convenience | Compared with written or telephone surveys, respondents need the least effort to finish the survey online; |
| Ease of data entry and analysis | The result of the survey can be easily saved and analysed in a digital format; |
| Question diversity | In order to generate references on roles for further study, various formats of questions should be asked to enrich the content. Therefore, open questions, multiple choice questions, ranking questions, and so on can easily be included in the online survey; |
| Low administration cost | Good for the researchers' financial constraints; |
| Ease of follow-up | It is easy to send a reminder to the respondents to finish the survey online; |

| | |
|---|---|
| Controlled sampling | Not applicable to the present study; |
| Large sample easy to obtain | With the global reach and convenience, the chances of a larger sample is increased; |
| Control of answer order | Not applicable to the present study; |
| Required completion of the survey | Only by finishing all the required fields can the respondents submit, which ensures the completion of the survey; |
| Go-to capabilities | Questions in the survey might be connected to each other, even though they are displayed in different session, so go-to functions make this easier to realize. |
| Knowledge of respondent vs. non-respondent characteristics | Not applicable to the present study; |

The insights attained from the online survey contribute the basic understanding of the roles of public libraries in the current Big Data wave. In order to reach a deeper comprehension, a semi-structured interview was next carried out with library directors on the feasibility and applicability of public libraries undertaking these roles. According to Dunn (2000, pp.79-80), there are three types of interviews for research: structured, unstructured, and semi-structured interview. Structured interviews follow the decided questions strictly. Unstructured interviews have no questions at all, which makes interviewees direct the interview all by themselves. Semi-structured interviews are in between of these two formats. The content of the questions is predetermined but they can be expanded and modified depending on the situation of the interview. Compared with the other two interview formats, semi-structured interviews work better when the informants come from different professional, educational, and cultural backgrounds. Moreover, the format is suited to exploring perceptions and opinions (Barriball & While, 1994). The semi-structured interview format was therefore selected for Study 1 to further interpret the survey findings.

### 3.2.1 The conduction of the online survey

Since no mature questionnaires could be used or referred to for the survey, a new questionnaire was designed. In order to include suitable content in the questionnaire, two exploratory steps were first carried out as the basis for designing the questionnaire:

- Step One: to summarize Big Data characteristics in public libraries
- Step Two: to attain possible name of roles

As discussed in Section 2.2, the characteristics of Big Data have been widely studied. Nevertheless, few studies have paid attention to Big Data characteristics in specific fields. Therefore, exploring Big Data characteristics specifically in the context of public libraries would not only be useful for the questionnaire design but also help to fill this research gap. The characteristics of Big Data are summarized from the perspective of data content, because data content would hardly change with any variance of data volume or data variety. Moreover, the content of data could reflect library services, making this perspective suitable to highlight library roles in the context of Big Data.

In Step One, reports or websites that contain public library statistics were carefully reviewed. The reason for choosing these websites and reports is that it could reflect what kind of data is generally stored and used within a public library. Therefore, the data content could be summarized. The samples were selected from Google with the keywords: "library statistics+country name". All European Union countries were selected for this information retrieval process. The reason for choosing countries from European Union was to limit differences in regions and policies. The list of European Union countries was retrieved from the Europa.eu website. In order to limit errors in the study, only websites or reports in English were finally used, because the language of the current study is English. In the end, only websites or reports in six countries (Bulgaria, Denmark, Finland, Ireland, Lithuania, and Malta) were selected. The description of the report or the key figures on the website were the main data to collect, since such descriptions and key figures should represent the main content of data. The result of data collection is presented in Table 3.

**Table 3: Data content of public libraries in EU countries**

| Country | Data description or key figures |
|---|---|
| Bulgaria | "The data refers to the number of libraries, library collection, readers, visitors, library collection loaned, employment and library staff. The survey included indicators for the number of books in libraries, continued editions (newspapers, magazines, bulletins and periodicals) and other documents in libraries (audio-visual and electronic documents, graphic and cartographic publications, patents and standards, microforms). The data indicate the total revenue, subsidies from the budget and the expenditure of libraries." (Republic of Bulgaria National Statistical Institute) |

| | |
|---|---|
| Denmark | Loan, stock, use of electronic resources (downloads), expenditure on material (Statistics Denmark) |
| Finland | Collections, Collections/inhabitants, disposals, acquisitions, acquisitions/1000 inhabitants, acquisitions/collections(100%), periodicals, E-material, Loans, customers, loans/inhabitants, lending circulation, interlibrary loans, expenditures, expenditures/inhabitants, economy, service points, openning hours, events and user taining, personnel (Kirjastot.fi) |
| Ireland | Income, expenditure, collection, issues, registered membership and charges, service points, agency services, service to schools (LGMA, 2012) |
| Lithuania | document stocks, number of users, loan of documents, number of visitors, professional stall, access to the Internet (Martynas Mažvydas National Library of Lithuania, 2015) |
| Malta | book acquisition, new library memebers, book loans, type of member, book loans by localities, Imports and exports of printed book material and periodicals in terms of value, Private final consumption expenditure in the domestic market (Natioanl Statistics Office) |

After reviewing statistics on the selected websites, the public library data content could be summarized as:

- Library resources. Data regarding collections, stock, staff, issues, service points, etc. This data provides a foundation for the libraries to offer services.
- Library economics. Data regarding expenditures, income, finances, or budgets. Such data indicates how well a library is operated financially.
- User information. Data regarding the number of users, the types of users, and so on. User groups will be clearly illustrated with the help of such data.
- Interaction between users and libraries. Data regarding use of library resources, loans, the number of visits, events, and user training. Such data shows how the libraries communicate with their users. Behaviors of users and libraries are also reflected by such data. For instance, visits are data describing how many people come to a library, loans imply individuals' preferences

for certain books, and the number of events organized by libraries represents the communication behaviors of libraries towards users.

These four aspects could be considered the content characteristics of Big Data in the public library. It can be concluded that data recorded in a public library would easily reveal the resource situation, financial situation, user information, and the interaction between users and libraries. Therefore, the design of the questionnaire should reflect library roles in handling data within these four areas.

Step Two is conducted by interviewing librarians in order to gain a basic understanding of the situation from a librarian's perspective and to glean hints and knowledge that could be useful for further expanding on library roles. Three librarians from a university library were interviewed. They had worked in the library environment for four, nine, and fifteen years respectively. University librarians were selected for this step in order to acquire a more general view and to learn from their experiences, as university libraries collaborate with researchers who have already worked with Big Data. These experiences might provide additional interpretations of Big Data in libraries, which would then enrich the discovery of library roles.

Based on these interviews, librarians do not have much experience with Big Data management in their daily work, and they only have a general understanding of Big Data. Nevertheless, they do hold a positive attitude towards applying Big Data in the library:

"That's the way to go in the future."

"It will come naturally."

Seven roles were mentioned during the interview: educator, marketer, data, data container, advocator, advisor, and developer. According to the study by Stejskal & Hajek (2015), public libraries not only serve citizens but also organizations. Therefore, an eighth role – organization server – was added to the survey. Based on the results of these two steps, a questionnaire was designed and then reviewed by other professionals. The items were rephrased and refined on the basis of their feedback. The target group of this survey is all Finnish librarians working in the public library.

The questionnaire has three parts, with a total of 28 questions. The first is used to collect demographic information (Q1-Q5). The target of the second part (Q6-Q11) is librarians' perception of Big Data, which echoes the current condition of Big Data in public libraries. Here, a four-V model of Big Data (Volume, Velocity, Variety, and Value) was employed, and four items were generated according to the content of each V. The average score of these four items is considered as the current perception of librarians on Big Data. The scale calculating these four items refers to the measurement put forward by

Schaufeli, Bakker, & Salanova (2006). The third part (Q12-Q27) concentrated mainly on realizing library roles. Q28 is an open question.

The questionnaire was delivered online and was officially sent out on January 27th, 2016. The valid time period for completing the questionnaire was from January 27th to March 10th, 2016. All the results were recorded automatically in a web-based format. The URL of the survey was sent to the reference email address of each public library in Finland. The email addresses could be found on the website Libraries.fi. In the end, 552 email addresses were collected, which was finished manually. Eventually, 49 responses were successfully attained. One was duplicated and thus deleted. In the end, 48 cases were analysed.

There were two reasons for this low feedback rate. Firstly, some librarians explained that they would not like to fill in the survey questionnaire owing to their insufficient knowledge of Big Data. Secondly, the survey invitation letter was sent to all these email addresses at the same time. For some libraries, emails with links sent to a massive list of recipients would be identified as spam and recorded in the spam folder, which decreased the chance for librarians to find the link to the online survey. Even though the 48 librarians who responded cannot be considered as fully representative of all Finnish librarians, their feedback was a valuable first-hand resource to further the study. Therefore, the survey results were analysed in order to provide a roadmap for future research. All the results were recorded and analysed with SPSS 17.0.

### 3.2.2 The conduction of the semi-structured interview

The interviewees were selected from libraries in big cities in Finland. The rationale for choosing librarians in big cities is that the city libraries of these cities have been pioneers in Finnish librarianship. They also have branch libraries in small or suburban areas. Therefore, the opinions of leaders in these big libraries could have a wide influence and represent a large purview. In the end, eleven interviews were carried out from October 26th to November 24th, 2016. The average time for each interview was 30 minutes. All these interviews were in English, and the audio of each interview was recorded and transcribed manually. During the interview, interviewees' daily data practices, their attitudes towards the eight roles examined in the survey, and their opinions on Big Data were discussed.

## 3.3 Research design of Study 2: Comprehending Big Data in librarianship

An analysis of Big Data definitions is conducted for Study 2. Therefore, a process to collect definitions is needed, which can be classified into two steps: literature collection and definition collection.

A cross-check method was employed to collect relevant literature. Articles from representative academic databases and journals were checked in order

to ensure the scope of the literature collection. The overall process of literature collection is shown in Figure 6:

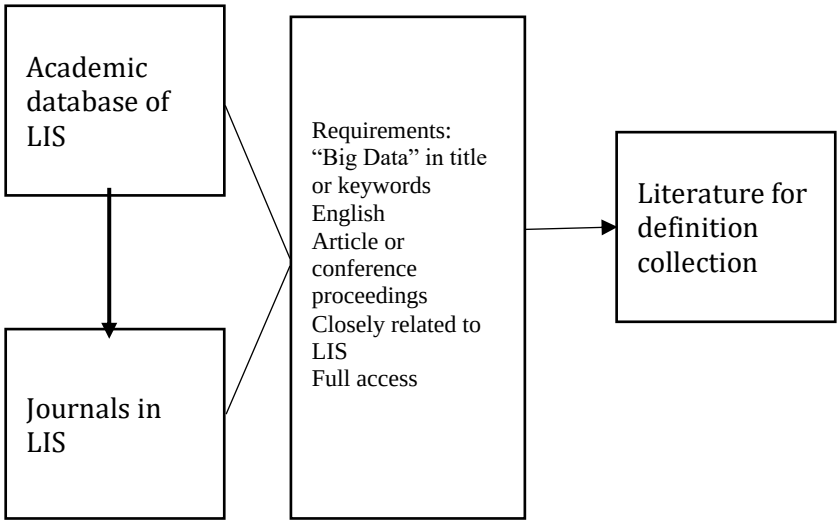



**Figure 6. The process of literature collection**

Two academic databases (EBSCO and Web of Science) were chosen as the main platforms for definition collection, with two considerations in mind. Firstly, they host a great number of scientific materials. Secondly, they have clear discipline search capabilities for LIS articles. The list of journals was retrieved from Scimago Journal & Country Ranking, which aims at ranking journals by indicators provided by Scopus. In the end, 179 English-language journals and 8 conference proceedings in the LIS domain were gained for cross-check. The name of each journal or conference proceeding was searched in both academic databases. If the name could not find in either database, articles from that journal or conference proceeding would be retrieved from other databases. In this way, the scope of searching relevant LIS literature could be guaranteed. "Big data" (with quotes) was used to search titles and keywords of articles. Moreover, the abstract of each article was reviewed so as to ensure that the major content of each article would be relevant to LIS. At this stage, 124 articles were identified.

Big Data definitions were then extracted from these 124 articles. According to Brown (1998, p.130), "definitions are declarative sentences which assert the matter of fact." In light of this definition of definition, it was decided to consider sentences with "big data is/represents/includes/consists of …", "big data refers to …", and "the definition of big data is …" as Big Data definitions. From this process, 35 definitions were collected. There are three origins: 1) generated by the author, 2) cited by the author, and 3) summarized from definitions from early studies by the author. These 35 definitions are used in a total of 39

articles, since in some cases authors directly quoted a definition from another article.

After collecting these 35 definitions, both statistical description and content analysis were conducted. The statistical description was accomplished with QDA Miner, which is an add-in function of Excel 2010. Word frequency was calculated, excluding introductory phrases like "big data is/represents/includes/consists of", "big data refers to", "the definition of big data is" that were used to locate Big Data definitions but would not affect the meaning of the definition through their exclusion. Moreover, such phrases would unnecessarily increase the frequency of certain words (such as the frequency of "big" and "data"). In addition, the term "data set" was replaced by the compound "dataset", since the meaning is identical but "data set" as two separate words could unnecessarily increase the frequency of "data" in the outcome. The content of each definition was also manually reviewed and summarized. Statistical description and content analysis reflect how Big Data is defined in LIS, and such reflection could thus contribute to the specific definition of Big Data in librarianship.

## 3.4 Research design of Study 3: Effectively organizing hashtags on Instagram

Study 3 examines effective ways of organizing hashtags so as to attain more "likes" and comments. The conduction of Study 3 will be explained in three sections: data collection (3.4.1), identifying hashtag locations (3.4.2), and model selection (3.4.3).

### 3.4.1 Data collection

Study 3 is intended to contribute to public libraries. Therefore, only library-related captions were considered. In order to collect enough captions meeting the requirement, #library was employed to collect a test dataset. This dataset was used to recognize all other hashtags relevant to the library content. The co-occurrence of hashtags with #library was calculated. In the end, #book, #read, #reading, and #bookstagram were also chosen to collect captions, owing to their high frequency and their high co-occurrence with #library. Moreover, the number of captions under each hashtag is more than 3 million, which ensures a great number of captions for collection. The collection was started on April 22nd, 2017, and ended on August 31st of the same year. PHP programming language was used to crawl the public content of Instagram, which can be accessed without an Instagram account. This means that the privacy of Instagram users is not violated by Study 3. The post ID, the post time of the caption generated, the number of "likes", the number of comments, and the caption were collected. The number of hashtags was calculated and recorded after collecting captions. The same caption would be collected more than once if the number of "likes" or

comments changed. Therefore, another column was added in the database to record the updating time of each caption, as is displayed in Table 4:

**Table 4: A sample of the database**

| Post ID | Post Time | No. of Comments | No. of Likes | Caption | No. of Hashtags | Updated Time |
|---|---|---|---|---|---|---|
| 1498198509****19723 | 4/22/17 0:00 | 12 | 237 | Tag a person who you think might wanna play with your balls!! ?Do You Want To Play With My Balls - $14.95 ? .#weirdshityoucanbuy #weirdshit #weird #book #booklover #bookclub #balls #ball #kids #forkids #funny #hilarious #lol #hi #lmao #lmfao #reading #reader #readers #literature #tagafriend | 20 | 2017/4/25 21:07 |
| 1552367366****43592 | 2017/7/5 17:43 | 0 | 41 | There's a good thing about the freedom of choice... #freedomofchoice #choice #bliss #positivity #bookstagram #instawriterscommunity #ilovestories #instawriterssociety #writersofinstagram #writersoninstagram #writers #ilovetowrite #instapoetry #instapoetrygram #lovetobeawriter # #lovetowrite #shewrites #loversofpoetry #wordporn #wordgasm #poemporns #spilledink #spilledinkpoetry #spilledinkcomp #wordsofwisdom #words #wordlover #wordlove #wordsmatter | 30 | 2017/7/7 7:54 |

At the first stage, 6,927,427 captions were collected that contain at least one of the five hashtags. After deleting duplications and captions not in English, 2,561,424 captions were saved for further analysis.

### 3.4.2 Identifying hashtag locations

After data collection, a method to calculate hashtag organization was designed. Since hashtag organization is defined as where and how the hashtags are located in a caption, the distribution and density of hashtags in a caption are necessary to know, which in Study 3 is considered as the hashtag location. Based on Figure 7, there are three main hashtag locations: in the front, in the middle and in the end.

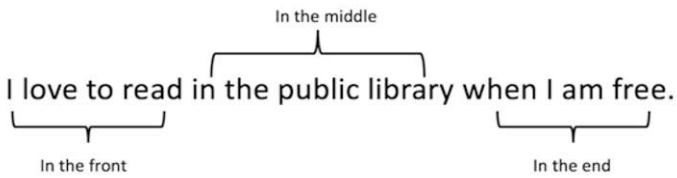



**Figure 7. Three main locations of hashtags**

To further identify hashtag locations, five variables are employed to reflect how hashtags are located:

- The index of the hashtags in a caption (HI), i.e., the order of hashtags appearing in a caption (for example, HI=[2,4,6] means the hashtags appear in the second, fourth and sixth place in the caption)

- The number of hashtags (HC)
- The number of words in a caption (WC), including the number of hashtags and common words
- Hashtag percentage (HP), calculated as HC/WC x 100%, which represents the density of hashtags in a caption
- Median (M), i.e., the median of the HI array
- Standardized deviation (SD), i.e., the standardized deviation value of the HI array

Both M and SD describe the distribution of hashtags in a caption. With these five variables and the three main locations, 15 locations are identified. The overall definitions and the explanation of each location are presented in Table 5:

**Table 5: The explanation of 15 locations**

| Name | AKA | Rule | Example |
|---|---|---|---|
| Only in the front | OF | One hashtag in the caption; in the front of the caption | I #love to read in the public library when I am free |
| Only in the middle | OM | One hashtag in the caption; in the middle of the caption | I love to read in the #public library when I am free |
| Only in the end | OE | One hashtag in the caption; in the end of the caption | I love to read in the public library when I am #free |
| All hashtags | AH | More than one hashtag in the caption; HP>=95% | #I #love #to #read #in #the #public #library #when #I #am #free |
| Most hashtags | MH | More than one hashtag in the caption; 80%<=HP<95% | I #love #to #read #in #the #public #library #when I #am #free |
| All at the beginning | AB | More than one hashtag in the caption; HP<80%; Hashtags are consecutively listed in the caption; captions starting with a hashtag | #I #love #to #read in the public library when I am free |
| All at the end | AE | More than one hashtag in the caption; HP<80%; Hashtags | I love to read in the public |

| | | | |
|---|---|---|---|
| | | are consecutively listed in the caption; captions ending with a hashtag | library when #I #am #free |
| All in the middle | AE | More than one hashtag in the caption; HP<80%; Hashtags are consecutively listed in the caption; captions neither started or ended with a hashtag | I love to read #in #the #public #library when I am free |
| Beginning and centralized at the end | BCE | More than one hashtag in the caption; HP<80%; Hashtags are not consecutively listed in the caption; both beginning and ending with a hashtag; M>2/3WC | #I love #to read in the public library when #I #am #free |
| Scattered | S | More than one hashtag in the caption; HP<80%; Hashtags are not consecutively listed in the caption; both beginning and ending with a hashtag; M<=2/3WC | #I love #to read in the #public library when I #am #free |
| Beginning and centralized in the front | BCF | More than one hashtag in the caption; HP<80%; Hashtags are not consecutively listed in the caption; beginning but not ending with a hashtag; SD<=0.3 | #I #love to #read #in #the public library when I am free |
| Scattered but not at the end | S not E | More than one hashtag in the caption; HP<80%; Hashtags are not consecutively listed in the caption; beginning but not ending with a hashtag; SD>0.3 | #I love #to read #in the public library when I am free |
| More at the end | ME | More than one hashtag in the caption; HP<80%; Hashtags are not consecutively listed in the caption; not beginning with a hashtag; M>2/3WC | I love #to read in the public #library #when #I #am free |
| Less in the front | LF | More than one hashtag in the caption; HP<80%; Hashtags are not consecutively listed in the caption; not beginning with a hashtag; M<=2/3WC; SD<=0.3 | I love to read #in #the #public #library #when I #am free |

| | | | |
|---|---|---|---|
| Scattered but not in the front | S not F | More than one hashtag in the caption; HP<80%; Hashtags are not consecutively listed in the caption; not beginning with a hashtag; M<=2/3WC; SD>0.3 | I love to read #in #the #public #library when I am #free |

### 3.4.3 Model selection

Regression models were created to outline patterns for effective hashtag organization. In order to avoid multicollinearity, three regression models were generated under three situations respectively:

1. HC=1
2. HC>1 and HP>=80%
3. HC>1 and HP<80%

According to Dranove (2009), the distribution of dependent variables decides the selection of the regression model. For Study 3, dependent variables are the number of "likes" and comments, so the distribution of these two dependent variables in three circumstances was visualized as shown below:

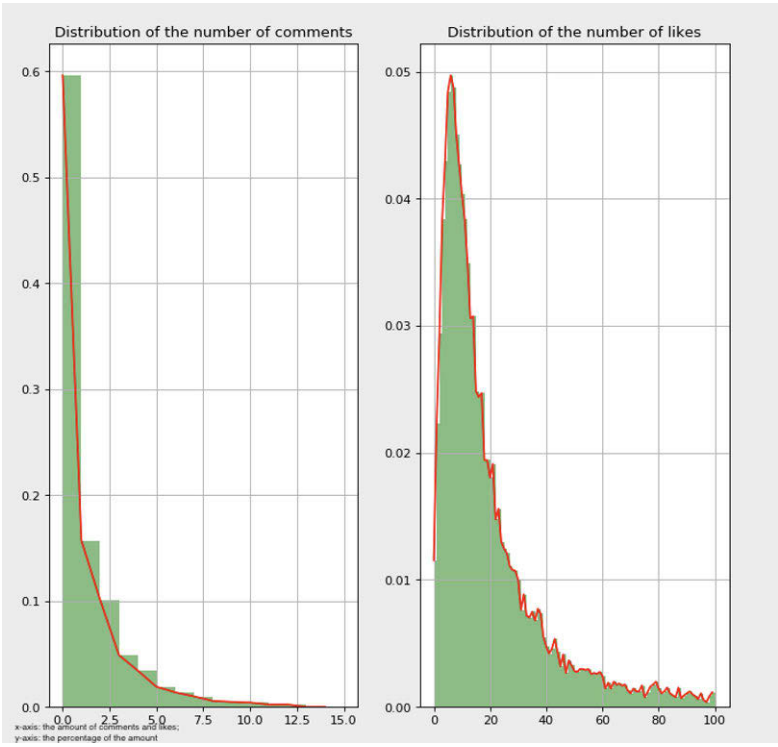



**Figure 8. The distribution of comments and "likes" with HC=1**

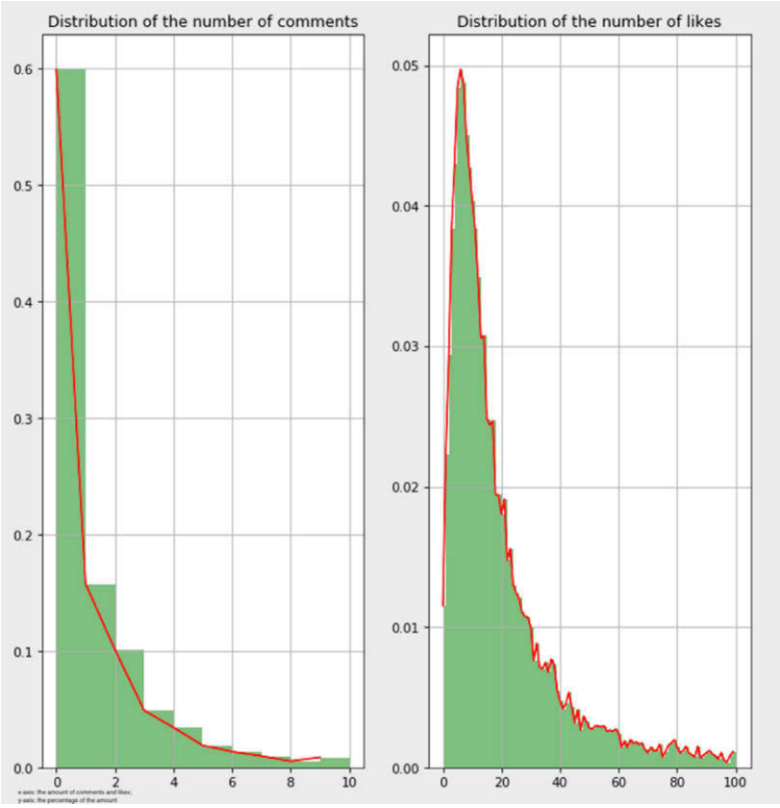



**Figure 9. The distribution of comments and "likes" with HC>1 and HP >=80%**

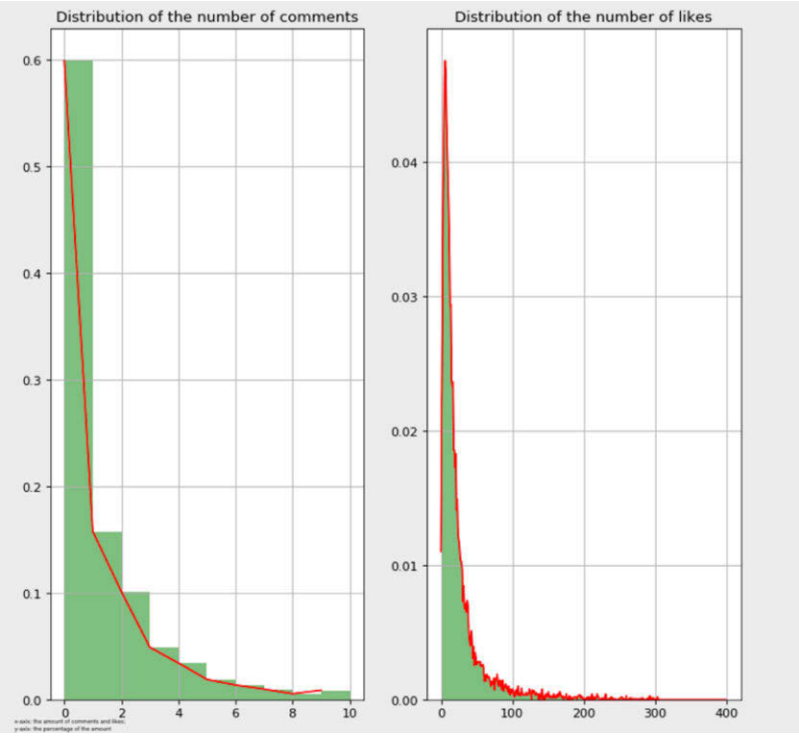



**Figure 10. The distribution of comments and "likes" with HC>1 and HP <80%**

In Figures 8–10, the left column presents the distribution of the number of comments, and the right column concerns the number of "likes". The X-axis in these three figures represents the number of comments or "likes". The Y-axis represents the percentage of captions with the corresponding number of comments or "likes". The distributions of both dependent variables in three circumstances are skewed to the right. However, the number of captions with no comments account for almost half of the overall captions, which makes the Poisson regression model suitable for the number of comments as a dependent variable, because the presence of excess zero observations and a long right tail is relevant to Poisson assumptions, based on the study by Famoye and Singh (2006).

As for the number of "likes", although the distributions in the three conditions share the same tendency as the number of comments, the zero observations do not count for a lot. Therefore, the Poisson regression model cannot be applied to the number of "likes". After checking the prerequisites of chi-square tests and linear regressions, it was concluded that the number of "likes" did not meet any of them. According to a study by Scott and Holt (1982), social studies use cluster or multistage sampling for economic reasons, which complicates the modification of the data. Thus, ordinary least squares (OLS) regression models could be beneficial, because OLS

parameter estimates are not seriously affected when data is subject to intracluster correlation. Since the variance of the number of "likes" under different locations is explored, each location could be one cluster. Therefore, ordinary least squares (OLS) regression was applied to forecast which hashtag location would cause more "likes".

## 3.5 Research design of Study 4: Outlining topics current readers like or dislike

Study 4 aims to highlight the topics current readers like and dislike. The conduction of Study 4 will be explained in three sections: data collection and pre-processing (3.5.1), the model creation for opinion polarity classification (3.5.2), and the model creation for emotion classification (3.5.3).

### 3.5.1 Data collection and pre-processing

In order to guarantee that reader-related content could be collected, captions with #read and #reading were downloaded from Instagram with PHP programming language. As in the previous study, all the captions were public content. At this stage, 4.6 million captions were collected. Text cleaning and pre-processing were conducted on these captions, which is illustrated by Figure 11:


**Figure 11. The process of data cleansing and pre-processing**

Duplicated captions and captions not written in English were excluded. Captions with quotes from books, movie lines, and so on were deleted as well, because such quotes would reveal the opinion and emotion of the characters or writers rather than the current readers. Special characters (e.g., #, @, +, &, %) were deleted from the text. Website links were also removed. Numerous hashtags are formed from combined words, and these were refined. For instance, "#agoodday" was expanded to "a good day". In Study 4, only 5,000 hashtags, picked out based on the ranking of hashtag frequency, were refined. The reason for picking 5,000 as the threshold was to ensure that only frequently used hashtags were included in the study. The cleaning

and pre-processing tasks were finished via Python 3. In the end, 2,035,706 captions were saved for further analysis.

### 3.5.2 The model for opinion polarity classification

For supervised machine learning, a training dataset with the truth (in this case, a dataset with labels of positive, neutral, and negative) is indispensable. However, no such dataset existed for Study 4. Lexical features were therefore extracted. Moreover, lexical features can avoid domain issues. This is illustrated in Figure 12:

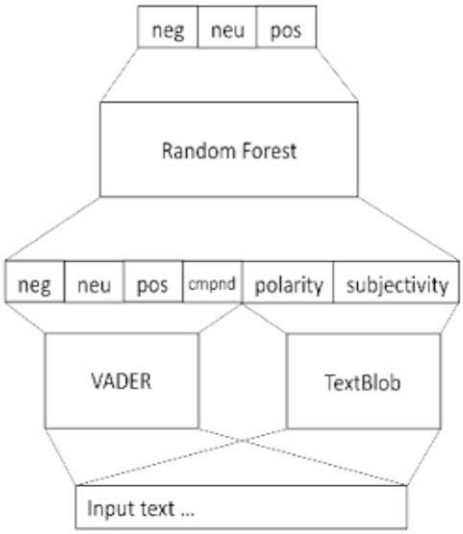



**Figure 12. Illustration of model creation for opinion polarity classification**

All the captions are represented by six features attained from VADER and Textblob, both of which are text mining packages in Python. The motivation for employing these two lexical classifiers is the absence of labels and issues of domain adaptation. A ready-to-use dataset containing tweets with opinion labels was used to create the model. The reason to choose this dataset is that both tweets and Instagram captions are generated on a social media platform. Even though domain issues between these two media cannot be ignored, this dataset has been used for sentiment analysis study (Gilbert, 2014), which indicates that the quality of the dataset is well ensured. Lexical features were extracted from the tweets as well. With the existence of features and truth, various supervised machine learning algorithms (naïve Bayes, SVM, decision tree, etc.) were applied to learn the truth from the training dataset, which in Study 4 is the tweets used by Gilbert for his

study. In the end, the model created by random forest achieved the best accuracy.

Verification was conducted in order to confirm the quality of this opinion polarity classification model. The model was applied to captions with clear opinions, which are reflected by delegate hashtags (see Table 6).

**Table 6: The result of testing the opinion polarity classification**

| Hashtag | Negative | Neutral | Positive |
|---|---|---|---|
| scary | 1580 | 641 | 361 |
| loser | 138 | 32 | 15 |
| upset | 134 | 11 | 12 |
| enjoyable | 2 | 14 | 60 |
| happy | 330 | 2200 | 42229 |
| fun | 181 | 4129 | 19457 |

It can be concluded that captions with these six hashtags have been classified properly. Most captions are labelled with the opinion polarity consistent with the meaning of the hashtag. Therefore, the sentiment analysis result generated by this model can be further employed.

### 3.5.3 The model for emotion classification

The emotions used in Study 4 are based on the study by Yadollahi, Shahraki, & Zaiane (2017): love, joy, sadness, fear, anger, and surprise. Each emotion belongs to one of the opinion polarities. As is listed in Table 7:

**Table 7: Six emotions and their antecedents**

| | Emotions | Antecedents |
|---|---|---|
| Positive | Love | The realisation of the target and the attainment of something people want or need. (These two antecedents are similar to those that create joy.) Sharing time and experience or communicating well with someone that the person believes is physically or psychologically attractive. |
| | Joy | Positive outcomes which initiate happiness, such as achieving success, or creating self-esteem. |
| Negative | Sadness | Already existing threats and undesirable outcomes. |
| | Fear | The interpretation of physical harm, loss, rejection, failure or situational factors that increase a person's perceived vulnerability |
| | Anger | Something that interferes with the execution of a person's plans or the attainment of their goals; something that creates physical or psychological pain. |
| Both | Surprise | A difference between a person's anticipation and subsequent experience. Positive surprise: delight at finding that your office window overlooks a garden; negative surprise: disappointed at finding that your window cannot be opened |

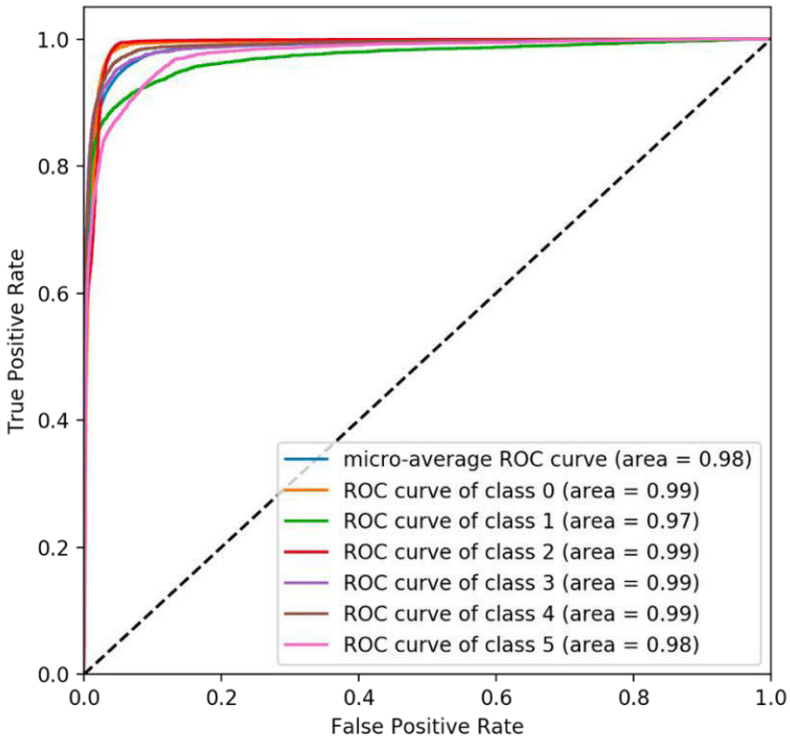
The antecedents of each emotion are explained based on the studies by Louis (1980) and Shaver, Schwartz, Kirson, & O'connor (1987). One thing should be emphasized, which is that surprise is an emotion that can belong to both positive and negative opinion polarities. The combination of the emotion and opinion polarity could deepen the discussion on readers' sentiment on current topics. A supervised machine learning model was generated for emotion classification with SVM algorithm, because the feasibility of SVM has been demonstrated for emotion classification by Joachims (1998). The training dataset is provided by Goel, Palaniappan, & Arora (2014), who are committed to conducting research to classify emotions. This dataset contains 89,832 pieces of data, and under each emotion, there are 14,972 reviews respectively. As a ready-to-use dataset, it has been pre-processed, with extra effort placed on feature collection to gain a valid classifier. The performances of the classifier are compared with the two feature selection methods (term frequency vs. tf-idf).

For term frequency, each review is transformed into an array that contains the frequency of each word in the 89,832 reviews. Tf-idf is short for term frequency-inverse document frequency, meaning the frequency of the term multiplied by inverse document frequency. The idf equation is written as:

$$\text{Equation 1: idf(t)} = \log \frac{1 + nd}{1 + df(d,t)} + 1$$

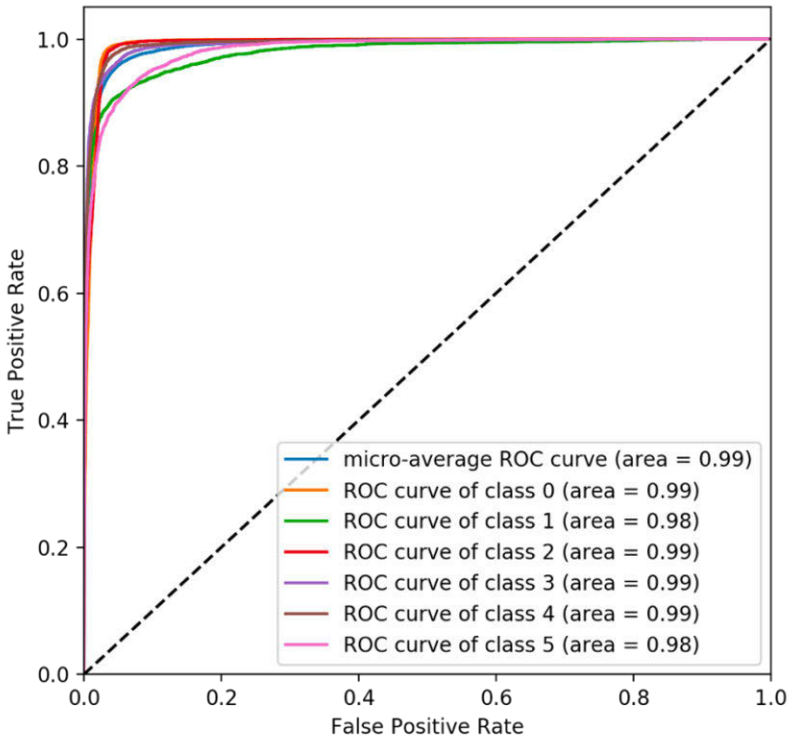
where *nd* is the total number of documents, and *df(d,t)* is the number of documents that contain the term *t*. Tf-idf reweights words that are rarely mentioned but which are more interesting (Pedregosa et al., 2011).

In order to compare the performance, the receiver operating characteristic (ROC) curves of classifiers established on these two features are visualized in Figure 13 and Figure 14. Hanley & McNeil (1982) state that the ROC curve is widely used to judge the discrimination ability of various methods for predictive purposes. The area under the curve corresponds to the possibility of it being correct.




***0:love, 1: joy, 2:suprise, 3:sad, 4:anger, 5:fear***

**Figure 13. The ROC curve of term frequency**




***0:love, 1: joy, 2:suprise, 3:sad, 4:anger, 5:fear***

**Figure 14. The ROC curve of Tf-idf**

As is shown in Figure 13 and Figure 14, when Tf-idf is applied for feature creation, the classifier achieves a slightly higher possibility (99%), which means the classifier works on this feature and can achieve more accurate results. Therefore, Tf-idf was chosen as the feature selection method for emotion mining. The generated classifier was applied to the target dataset without further consideration of the domain adaptation. Even though this training dataset is not generated based on Instagram data, the 99% possibility of identifying the correct emotion demonstrates the capability of this classifier.

# 4. Results

As explained in Section 1.1, the overall purpose of this doctoral study is split into three subordinate aims. Four studies are conducted to achieve these three aims thus to finally accomplish the overall purpose of this doctoral study. In this section, the results of the four studies outlined above are presented. Firstly, the roles that (Finnish) public libraries should undertake in the context of Big Data are presented. Secondly, a summarized definition of Big Data specifically for librarianship is put forward on the basis of content analysis and statistical description of Big Data definitions appeared in former LIS studies. Lastly, the results of the two studies analyzing hashtags generated on Instagram for public libraries are presented.

## 4.1 Results of Study 1 : Outlining roles of public libraries in the context of Big Data

Study 1 is composed of an online survey and semi-structured interviews, the study findings of which will be presented accordingly:

### 4.1.1 Results of the survey

There are 48 respondents, two-thirds of whom were female. Most were between 26 and 40 years old. More than 80 percent of the respondents have bachelor or higher degrees. Moreover, 81 percent of the respondents have been working in the Finnish public library for more than three years, and most of them undertake more than one responsibility. Such an educational and professional background implies that they are qualified to comprehend and interpret Big Data in librarianship. Furthermore, according to Figure 15, librarians who have no hands-on experience with Big Data do not greatly outnumber than those who have some experience, despite the infancy of Big Data in the field.

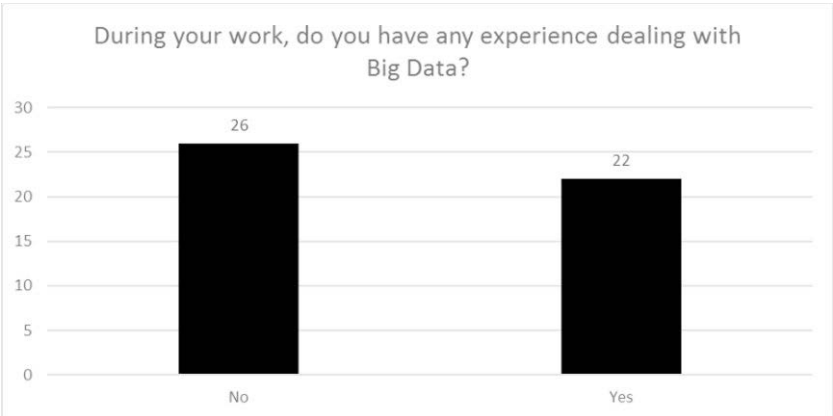



**Figure 15. Responses on experience with Big Data**

The questionnaire was designed with Likert 7, whereby 1 means totally disagree and 7 means totally agree. As seen in Table 8 (the Item column records the serial number of statements regarding different roles), the mean values of each role range from 4.10 to 5.33, which means no strong opinions on these roles are received via the survey. To some extent, responses like "the concept of Big Data is rather vague" and "not sure what Big Data is" could explain such slight agreement or disagreement on the roles. However, respondents tend to agree slightly on the provided roles. For statement 17 ("The public library should establish [a] data warehouse to store and preserve data generated from users or from projects incorporating other public libraries"), the average score is lower than 4, which indicates a slight disagreement. Respondents agree most with statement 20 ("The public library should give users tools or instructions rather than the actual result when users have trouble in managing personal information").

**Table 8: Opinions on different roles**

| Roles | Item | Mean value | Overall mean value |
|---|---|---|---|
| Marketer | I12 | 5.27 | 5.06 |
| | I13 | 4.85 | |
| Educator | I14 | 4.83 | 5.22 |
| | I15 | 5.60 | |
| Data Organizer | I16 | 4.54 | 4.62 |
| | I18 | 4.69 | |
| Data Container | I16 | 4.54 | 4.64 |
| | I17 | ***3.81*** | |
| | I19 | 5.58 | |
| Advocator | I20 | ***5.79*** | 5.33 |
| | I21 | 4.86 | |
| Advisor | I22 | 5.38 | 4.96 |
| | I23 | 4.54 | |
| Developer | I24 | 5.00 | 4.86 |
| | I25 | 4.71 | |
| Organization server | I26 | 4.10 | 4.10 |
| | I27 | 4.10 | |

When it comes to individual opinions, 28 people agree on the listed roles, since their average score on these roles is higher than 4.5. The lowest individual average score is 2.8 and the highest is 6.8. At the end of the survey, an open question was asked to add additional roles. However, no additional roles were mentioned. In conclusion, the survey findings reveal that opinions on these eight roles are stronger than neutral. However, librarians do not have a full comprehension of Big Data.

### 4.1.2 Result of the interview

The interview centers on exploring the possibilities of Big Data and the possible roles of Big Data in the public library. Even though Big Data is new to the interviewed library directors, they hold optimistic opinions:

> "I am very positive about it. I think using Big Data will give better tools to serve citizens, to show politicians what we are doing…"
> "There are lot of things where we can dig into when we learn Big Data…"
> "Well, there are risks as we told. But it could be a huge possibility to us to use that [Big Data]…"

They consider Big Data a good resource for understanding patrons, making better decisions, and advocating data reusability, which means the reuse of data generated not only within the library but also in other public sectors. One example was offered during the interview: drive-in services. Public libraries could use the parking information around the library building. Such information could help public libraries develop services to guide users where to park. If the library building is located in a busy area, a drive-in service could be developed to simplify the process of borrowing and returning books, which in turn boosts the use of the library. In addition, half of the interviewees recognized the significance of understanding the needs of society before providing services derived from Big Data. However, their optimistic opinions are established on three preconditions: employees who have knowledge concerning Big Data utilization, resources that could be allocated to build the infrastructure for Big Data utilization, and authorization that guarantees the protection of personally sensitive information.

Interviewees were asked to discuss their opinions on the eight roles considered in the survey. Even though they do not have experience on handling Big Data, they connected these roles to projects they had done and put forward their assumptions. All in all, they agree on the eight roles, owing to the scope of these roles covering the main responsibility of the public library. However, they highly recommend that the role of data container be undertaken by one or a few big municipality libraries. The National Library of Finland was referred to as functioning as a data container, since one of the responsibilities of the National Library of Finland is to provide data services. The municipality libraries could undertake a such role as well to serve other small public libraries in the area. In addition, since the library system in public libraries in Northern Finland was provided by a vendor, it could be a

solution to outsource such a role to professional companies rather than assign it to the library.

The role of data organizer is accepted by the interviewees, because public libraries in different regions could have different strategies, which leads to different requirements for data. Nonetheless, such a data organizer role should not be necessary for each public library, because small public libraries are led by the main library in the region. The main library can serve in this role and provide services for the subordinate libraries.

The roles of educator, marketer, adviser, and advocator are encouraged by the interviewees, because their libraries undertook the same roles to introduce social media to citizens. They anticipate the same responsibilities with the advent of Big Data. However, all the interviewees stated that they have limited resources to act in these roles. As such, a new role was proposed: facilitator.

> […] the way we could cover educator, marketer, adviser and advocator is if the libraries work as a facilitator […] so the library again is the medium between the actual experts of Big Data […] and provide the possibilities, the means to meet people and help them meet the actual experts.

When it comes to developer and organization server, interviewees are positive as well. They think libraries are in a good position to develop tools or services with cutting-edge technologies. Therefore, using Big Data to develop new tools should not be excluded either. In addition, interviewees mentioned that libraries are looking to broaden their service scope. Therefore, it is a feasible idea to act as organization server. On one hand, such a role helps libraries to serve organizations, which broadens their service scope. On the other hand, it could strengthen the collaboration between the library and other organizations. However, limited resources would hinder the realization of this role.

## 4.2 Result of Study 2: Comprehending Big Data in librarianship

In Study 2, thirty-five definitions were analysed to define Big Data. Content analysis and statistical description of these definitions were conducted so as to abstract the key aspects of Big Data definitions in LIS research. There are two definitions (D9 and D32) not included in the further analysis process because both of them are derived from previous studies and simply abstract the main perspectives of Big Data concepts (D9: a product-oriented, process-oriented, cognition-oriented, and social movement perspective; D32: technical, historical, and ideological perspective). However, their formalities

to discuss Big Data perspectives lay the foundation of the study to conduct the analysis.

Five aspects were summarized to present the analysis result: the question of what Big Data refers to (4.2.1), the characteristics of Big Data (4.2.2), and the requirements, challenges, and benefits of Big Data (4.2.3). Analysis of word frequency and definition similarity were also conducted (4.2.4).

### 4.2.1 What does Big Data refer to?

After analyzing the content of the collected definitions, two orientations can be broadly identified: ability-oriented and data-oriented. In ability-oriented definitions, Big Data is considered the technology, process, or tools to manipulate large amounts of data. In data-oriented definitions, Big Data is recognized as either data or information. It should be noted that data and information are not differentiated in the study in question, given that information can be considered data that is refined for further use (Bellinger, Castro, & Mills, 2004, p.1). D29 covers both orientations.

### 4.2.2 The characteristics of Big Data

There are thirty-one definitions that address the characteristics of Big Data. After reviewing the definitions, four characters can be identified which corresponds to four V's: volume, velocity, variety and veracity (Buhl et al., 2013). The summary of characteristics covered by each definition is listed in Table 9:

**Table 9: Summary of Big Data's characteristics in the LIS definitions**

| NO. of Def. | Volume | Velocity | Variety | Veracity |
|---|---|---|---|---|
| D1 | ★ | | | |
| D2 | ★ | ★ | ★ | |
| D3 | ★ | ★ | ★ | ★ |
| D4 | ★ | | ★ | |
| D5 | ★ | | | |
| D6 | ★ | | ★ | ★ |
| D7 | ★ | ★ | ★ | |
| D8 | ★ | | ★ | ★ |
| D10 | ★ | | | ★ |
| D11 | | ★ | | ★ |
| D12 | ★ | | | ★ |
| D13 | ★ | ★ | | ★ |
| D14 | ★ | | | |
| D15 | ★ | | | |
| D16 | | | | |

| | | | | |
|---|---|---|---|---|
| D17 | ★ | | | |
| D18 | ★ | | | ★ |
| D19 | ★ | | | |
| D20 | ★ | | | ★ |
| D21 | | | | |
| D22 | ★ | | ★ | |
| D23 | | ★ | ★ | ★ |
| D24 | ★ | ★ | ★ | |
| D25 | ★ | ★ | ★ | |
| D26 | ★ | | | |
| D27 | ★ | | | |
| D28 | ★ | ★ | ★ | ★ |
| D29 | ★ | | | ★ |
| D30 | ★ | | | ★ |
| D31 | ★ | ★ | | |
| D33 | ★ | ★ | ★ | |
| D34 | ★ | | | ★ |
| D35 | ★ | ★ | | |
| | 29 | 12 | 12 | 14 |

Two definitions (D3 and D28) cover all four characteristics. The totals of each characteristic in the definitions are 29, 12, 12, and 14, which indicates that volume is much more emphasized in the analysed definitions than the other three V's (velocity, variety, and veracity).

### 4.2.3 Challenges, demands, and benefits

Ten definitions mention the challenge of Big Data. Eight definitions include the demand to cope with Big Data. Both challenge and demand are proposed from the technical point of view. The benefits of Big Data are also acknowledged by eight definitions, which are represented by useful insights and patterns, new values, and solutions to previously unsolved issues. It can be concluded that the realization of Big Data benefits leads to the process of value creation, which is another V feature (value) of Big Data (Storey & Song, 2017).

### 4.2.4 Word frequency and similarity analysis

In addition to the summarization of five aspects of Big Data definitions in LIS literature, the word frequency of the collected definitions was calculated.

**Table 10: Word frequency of analysed definitions**

| Word | Frequency | Number of |
|---|---|---|
| Data | 50 | 24 |

| | | |
|---|---|---|
| Large | 20 | 18 |
| Information | 10 | 7 |
| Big | 9 | 6 |
| Velocity | 9 | 9 |
| Volume | 9 | 8 |
| High | 8 | 6 |
| Variety | 8 | 8 |
| Analysis | 7 | 7 |
| Database | 7 | 6 |
| Datasets | 7 | 6 |
| Process | 7 | 7 |
| Sets | 7 | 5 |
| Technologies | 7 | 7 |
| Tools | 7 | 7 |
| Complex | 6 | 6 |
| Size | 6 | 5 |
| Storage | 6 | 6 |
| Systems | 6 | 4 |
| Techniques | 6 | 5 |
| Traditional | 6 | 6 |
| Processing | 5 | 5 |
| Volumes | 5 | 5 |

As presented in Table 10, only words mentioned at least five times are listed, in order to focus only on the more frequently used words. The frequency of "data" is much higher than the other listed words. This can be attributed to the fact that twenty-two definitions refer to Big Data as data. Moreover, "information" is the third-most mentioned word. Both words are relevant to the question of what Big Data refers to. The four V's summarized in Table 9 are all on the list. The list also includes "high", "size", "complex", "large", and "big", which are relevant to one of the four V's. This corresponds to the aspect of the characteristics of Big Data. "Analysis", "process", "technologies", "tools", "storage", "techniques", "traditional", and "processing" are semantically related to the aspects of challenge and demand. Therefore, it can be concluded that four aspects are reflected through the word frequency calculation.

A similarity analysis of the definitions was also conducted. The similarity analysis is based on the word used in each definition. The more words that two definitions share in common, the more similar they are. As is visualized in Figure 16:

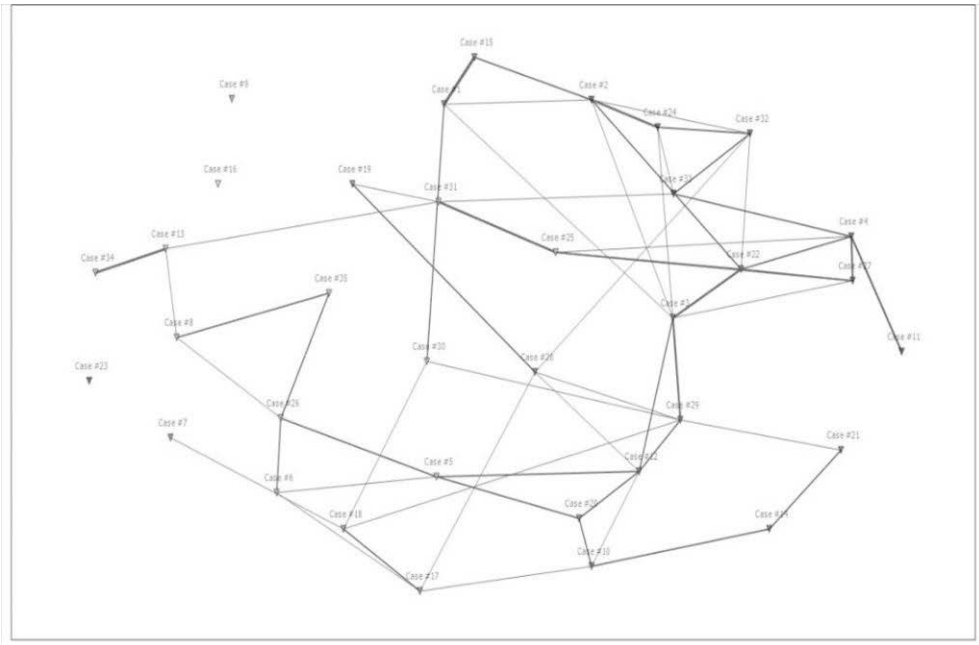



**Figure 16. The network of definition similarity**

Only definitions with a match of at least 50 percent are linked in Figure 16. Each node exhibits as each definition. The shorter the distance between two nodes, the more similar these two definitions are. Three definitions (D9, D16, D23) are not linked to any definition, which reveals the unique content of each definition. D3 is linked the most (seven links with other definitions), implying that D3 could be a definition agreed upon by most LIS scholars. D22, D29, and D33 are all linked with 6 other definitions, with the same implication of centrality as D3. The results of the definition similarity analysis could help to group related definitions and thus to compile the definition of Big Data in the context of librarianship in combination with the results of word frequency and content analysis.

Based on all the above, Study 2 defines Big Data as data with a large size, fast-growing speed, and various types, which can complicate data handling techniques but also boost the creation of technological solutions. Value is generated by the proper operation and use of Big Data.

## 4.3 Result of Study 3: Effectively organizing hashtags on Instagram

For Study 3, hashtag organization is represented by the locations of hashtags. The findings of Study 3 are presented in two sub-sections: the statistical description of variables (4.3.1) and the result of regression analysis (4.3.2).

### 4.3.1 The statistical description of variables

There are 15 types of location identified. The number of captions in each location is shown in Figure 17:

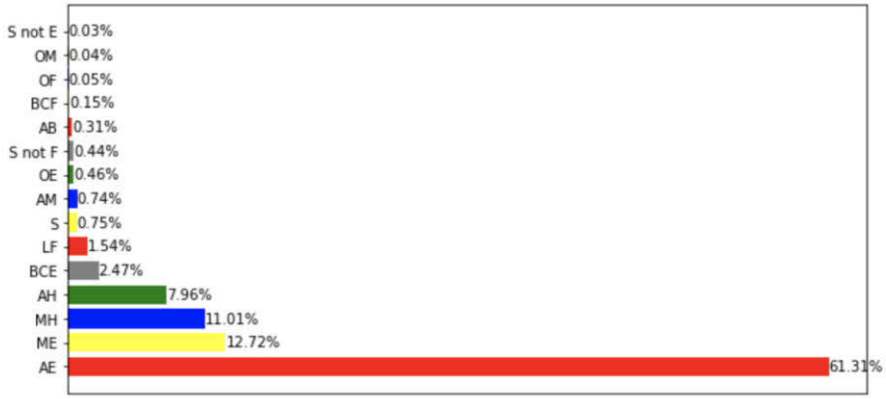



**Figure 17. The percentage of captions per hashtag location**

Figure 17 visualizes the disproportionate distribution of captions in each hashtag location. Captions in location scattered (S), all in the middle (AM), only in the end (OE), scattered but not in the front (S not F), all in the beginning (AB), beginning and centralized in the front (BCF), only in the front (OF), only in the middle (OM), and scattered but not at the end (S not E) each account for less than 1 percent. More than 60 percent of captions are in the location all in the end (AE). Considering that 12.72% of captions belong to the location more in the end (ME), there are 74.03% captions with hashtags mainly used at the end. This implies that library-related captions on Instagram tend to put hashtags at the end.

The aim of Study 3 is to achieve solutions for organizing hashtags effectively for public library-related captions on Instagram. In this study, the effectiveness of hashtag organization is reflected by the number of comments and "likes". Apart from hashtag location, there are other variables needed for the analysis:

- The number of hashtags (HC): the number of hashtags each caption contains
- The number of words (WC): the number of words each caption contains
- Time difference (TD): the time difference between the time when the number of comments and "likes" of a caption was last updated in the database and the time when the caption was initially created; TD is measured by the hour
- The number of comments (C): the number of comments each caption contains at the time when the database is last updated
- The number of "likes" (L): the number of "likes" each caption contains at the time when the database is last updated

**Table 11: Statistical description of variables**

| | HC: No. of hashtags in a caption | WC: No. of words in a caption | No. of comments (C) | No. of likes (L) | Time difference (TD) |
|---|---|---|---|---|---|
| *M* | 16 | 46 | 2 | 70 | 58.97 |
| *StD* | 8.71 | 45.03 | 15.34 | 626.58 | 66.10 |
| Mode | 29 | 30 | 0 | 14 | |
| Min | 1 | 2 | 0 | 0 | 2 |
| 25% | 8 | 20 | 0 | 17 | 21.22 |
| 50% | 15 | 32 | 1 | 30 | 37.74 |
| 75% | 23 | 53 | 2 | 54 | 68.72 |
| Max | 248 | 445 | 12,648 | 736,842 | 429.83 |

The statistical description of these variables is presented in Table 11. According to Table 11, most captions contain around 20 hashtags and 50 words, and were finally updated within three days. The mean value of the number of "likes" is 30, but the mean value of the number of comments is only one. Furthermore, the mode and the 25th percentile of the number of comments are both zero. Actually, there are 1,269,624 out of 2,561,424 captions with no comments by the last updating time. The distribution of the number of comments is shown in Figure 18:

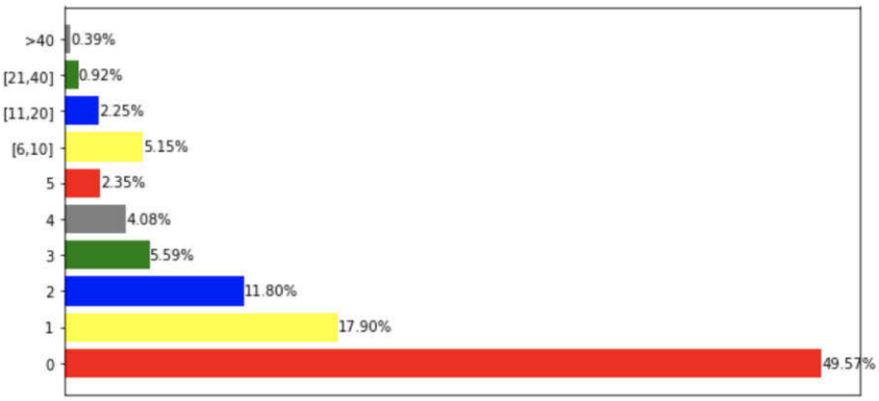



**Figure 18. Bar chart of the number of comments**

As Figure 18 illustrates, almost half of the collected captions have no comments, and only 15% of captions have at least 4 comments. Nonetheless, nearly 2% of captions have less than 2 "likes". Such a finding indicates that people more tend to like a caption rather than comment on the caption within a certain amount of time.

Since nearly half of captions have no comment, the proportion of zero comment captions with each location type was explored. Such exploration could be helpful to identify locations that are more likely to gain no comments. Hypothetically, the percentage of zero comment captions within

each location type should be similar to the overall percentage: 49.57%. If that percentage is higher in a certain location, it could imply that this location is more relevant to zero comment than other locations. Fisher's exact test (Bower, 2003) was conducted to verify the hypothesis.

**Table 12: The result of Fisher's exact test**

| Location abbrev. | Percentage | p-value of FE test |
|---|---|---|
| AE | 49.68% | 1.00 |
| MH | 49.83% | 1.00 |
| ME | 42.46% | 0.39 |
| AH | 58.87% | 0.25 |
| BCE | 50.22% | 1.00 |
| LF | 48.48% | 1.00 |
| S | 56.48% | 0.39 |
| AM | 46.82% | 0.78 |
| **AB** | **63.50%** | **0.06** |
| S not F | 52.53% | 0.78 |
| OE | 59.07% | 0.25 |
| BCF | 58.07% | 0.31 |
| S not E | 56.87% | 0.39 |
| OF | 53.64% | 0.67 |
| OM | 52.63% | 0.78 |

As listed in Table 12, thirteen of fifteen hashtag locations have a percentage of zero comment captions that is higher than the overall percentage. However, only the percentage of all in the beginning (AB) has a p-value smaller than 0.1, which means that only the greater percentage value of location AB is significantly higher. This finding indicates that more zero comment captions can be found when a public library-related caption puts all hashtags at the beginning of the caption.

### 4.3.2 Result of regression analysis

After deciding the regression model for these two dependent variables, outliers of the number of comments and "likes" were excluded. In this study, the 98$^{th}$ percentile was chosen as the threshold. Thus, the value higher than the 98th percentile of each variable was deleted.

The results of Poisson regression models applied to the number of comments in three conditions are shown in Table 13:

**Table 13: Poisson regression results**

| HC = 1 | | | HC > 1 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | HP > = 80% | | | HP < 80% | | |
| | **Coef** | **p_value** | | **Coef** | **p_value** | | **Coef** | **p_value** |
| Intercept | **−0.4230** | **0.0000** | Intercept | **−0.6136** | **0.0000** | Intercept | **−0.6289** | **0.0000** |
| OF | 0.1329 | 0.1250 | MH | **0.1877** | **0.0000** | AE | **0.2014** | **0.0060** |
| OM | **0.1996** | **0.0000** | TD | 0.0011 | 0.0050 | AM | **0.2169** | **0.0010** |
| TD | 0.0007 | 0.1840 | L | **0.0088** | **0.0000** | BCE | 0.1081 | 0.0840 |
| WC | **0.0035** | **0.0000** | | | | BCF | 0.0083 | 0.9030 |
| L | **0.0093** | **0.0000** | | | | LF | **0.1885** | **0.0050** |
| | | | | | | ME | **0.2942** | **0.0000** |
| | | | | | | S | 0.1072 | 0.1030 |
| | | | | | | S not E | 0.0515 | 0.5950 |
| | | | | | | S not F | **0.2476** | **0.0020** |
| | | | | | | TD | 0.0004 | 0.0980 |
| | | | | | | HC | **0.0054** | **0.0000** |
| | | | | | | WC | **0.0022** | **0.0000** |
| | | | | | | L | **0.0084** | **0.0000** |
| **No. of observations** | | 13,673 | **No. of observations** | | 461,723 | **No. of observations** | | 1,905,230 |
| **Scale** | 1 | | | | | | | |

As is shown in Table 13, the value of scale equals one, which means a good performance of Poisson regression. Time difference is not a significant variable influencing the number of comments in either situation. To put it another way, when the other variables are certain, increasing time difference will not result in more comments. The number of "likes" is positively significant to the number of comments in all situations. Certain hashtag locations in each situation can contribute to increasing the number

of comments, which is much higher than other significant variables. The detailed explanations of each condition are:

- HC=1. Although word count and the number of "likes" are both positively relevant to increase the number of comments, their influences are much lower than putting the hashtag in the middle of the caption. That is to say, placing the hashtag in the middle is the best solution to boost comments considering the given variables. Furthermore, placing the hashtag at the end has a negative influence on the number of comments, which implies such a location should be avoided when there is only one hashtag used in public library-related captions.
- HC>1 and HP>=80%. In this situation, hashtag count and word count are not included owing to their little variance from each other. Similar to the aforementioned situation, placement of hashtags is more effective than increasing the number of "likes" in terms of increasing the number of comments. In addition, creating a caption with more than 80% but less than 95% hashtags is better than creating an all-hashtag caption from the perspective of attaining comments.
- HC>1 and HP<80%. All at the beginning is the worst location to place hashtags compared with the other nine locations, which can cause a decrease in comments. However, only locations as AE, AM, LF, ME, and S not F have significant p-value. According to the rules to define each location in Table 5, all of these significant locations include fewer hashtags in the front of the caption. Therefore, not placing hashtags in the front of captions is recommended when more than one hashtag can be used, and the hashtag percentage of the caption is less than 80%.

The results of OLS models applied to the number of "likes" in three conditions are shown in Table 14:

**Table 14: OLS regression results**

| **HC = 1** | | | **HC > 1** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **HP > = 80%** | | | **HP < 80%** | | |
| | **Coef** | **p_value** | | **Coef** | **p_value** | | **Coef** | **p_value** |
| Intercept | **14.4684** | **0.0000** | Intercept | **27.0755** | **0.0000** | Intercept | **4.2364** | **0.0000** |
| OF | **−2.4983** | **0.0090** | MH | **7.0446** | **0.0000** | AE | **5.4672** | **0.0000** |
| OM | **−2.1049** | **0.0490** | TD | 0.0246 | 0.0050 | AM | **5.2808** | **0.0000** |
| TD | **0.0104** | **0.0340** | **C** | **7.4925** | **0.0000** | BCE | **6.8520** | **0.0000** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| WC | **0.0990** | **0.0000** | | | BCF | 0.0318 | 0.9630 |
| C | **7.8847** | **0.0000** | | | LF | **1.6408** | **0.0000** |
| | | | | | ME | **6.6860** | **0.0000** |
| | | | | | S | 0.2190 | 0.6360 |
| | | | | | S not E | 1.9587 | 0.1620 |
| | | | | | S not F | 0.4595 | 0.3960 |
| | | | | | TD | **0.0190** | **0.0000** |
| | | | | | HC | **1.2241** | **0.0000** |
| | | | | | WC | **0.0376** | **0.0000** |
| | | | | | C | **7.4770** | **0.0000** |
| **No. of observations** | 13,673 | | **No. of observations** | 461,723 | **No. of observations** | 1,905,230 | |
| **Ajusted $R^2$** | 19.20% | | **Ajusted $R^2$** | 13.40% | **Ajusted $R^2$** | 24% | |

The value of adjusted $R^2$ varies from 13.4% to 24% in three situations, which means the selected variables could explain approximately 20% of the variance of the number of "likes". According to the studies by Geurin-Eagleman & Burch (2016), Hu et al. (2014), and Stuart et al. (2017), factors such as the content of the image posted on Instagram and the demographic situation of the Instagram account could lead to the change of the number of "likes", which are not considered in this study. Therefore, a moderated value of adjusted $R^2$ is expected. Since this study highlights the organization of hashtags, included variables in the OLS regression model are limited. Therefore, such moderate adjusted $R^2$ values are accepted. Certain location types contribute most of the variance of the number of "likes". However, the number of comments has a stronger positive influence on the number of "likes" than the other way around. The detailed explanations of these three conditions are:

- HC=1. To place the hashtag in the front or the middle of the caption has a negative influence on the number of "likes". Time difference is positively significant here. That is to say, when

other variables are certain, the longer after the post has been created on Instagram, the more "likes" that post will obtain. The number of words in the caption has a significant influence as well. However, the increment caused by time difference or the number of words is not as strong as the change caused by placing the hashtag in different locations.

- HC>1 and HP>=80%. The location most hashtags is still a better location than all hashtags, which is the same for increasing the number of comments.
- HC>1 and HP<80%. Under such circumstances, the number of comments could positively change a greater number of "likes" compared with other variables. However, certain hashtag locations work better than time difference, hashtag count, and word count. All other locations work better than placing hashtags all in the beginning (AB). However, only AE, AM, BCE, LF, and ME are significantly better than AB. These five significant locations can also be summarized as a location having fewer hashtags in the front.

The result of regression analysis indicates that certain hashtag locations work better in attaining comments and "likes" for the caption, which could suggest ways of effectively organizing hashtags to enhance the communication function.

## 4.4 Result of Study 4: Outlining topics current readers like or dislike

In order to understand current readers from a sentiment point of view, Study 4 combines two methods of analysis: opinion polarity classification (4.4.1) and emotion classification (4.4.2).

### 4.4.1 Result of opinion polarity classification

The number of captions in each opinion polarity is shown in Figure 19:

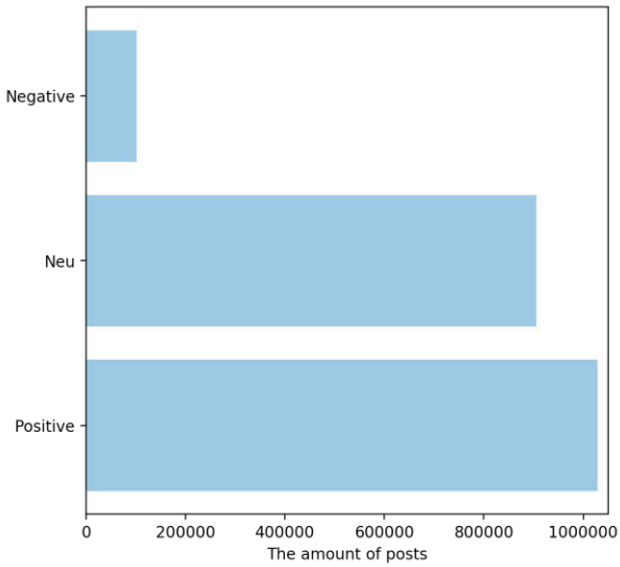



**Figure 19. The number of posts in each polarity**

As seen in Figure 19, there are far more positive and neutral posts than negative posts, which demonstrates that captions with #read and #reading do not incline to be negative.

In order to advance the analysis, representative topics in each opinion polarity were outlined. Representative topics can be identified by the frequency of hashtags in each polarity, since hashtags function to denote the topics of the caption on Instagram (Giannoulakis & Tsapatsoulis, 2015; Ma et al., 2013). Fifty was chosen as the threshold to select representative topics. That is to say, only the top 50 hashtags were chosen in each polarity, as listed in Table 15:

**Table 15: Top 50 hashtags in each polarity**

| | all | pos | neu | neg |
|---|---|---|---|---|
| hashtags in all groups (30) | books | books | books | books |
| | bookstagram | bookstagram | bookstagram | bookstagram |
| | book | book | book | book |
| | bookworm | bookworm | bookworm | bookworm |
| | booklover | love | booklover | booklover |
| | love | booklover | love | bibliophile |
| | bibliophile | reader | bibliophile | bookish |
| | bookish | bookish | bookish | booknerd |
| | reader | bibliophile | reader | reader |
| | booknerd | booknerd | booknerd | bookaddict |
| | bookaddict | bookaddict | bookaddict | instabook |
| | literature | literature | literature | writing |
| | library | library | library | library |
| | instabook | words | instabook | bookstagrammer |
| | words | author | words | bookphotography |
| | author | instabook | author | igreads |
| | bookphotography | bookaholic | bookphotography | bookaholic |
| | bookaholic | bookphotography | bookaholic | literature |
| | igreads | igreads | igreads | writer |
| | bookstagrammer | bookstagrammer | bookstagrammer | art |
| | writing | writer | writing | booksofinstagram |
| | writer | poetry | writer | bookshelf |
| | poetry | writing | poetry | poetry |
| | art | life | art | quotes |
| | bookshelf | quotes | bookshelf | love |
| | quotes | bookshelf | quotes | writersofinstagram |
| | booksofinstagram | art | booksofinstagram | words |
| | life | booksofinstagram | life | instabooks |
| | writersofinstagram | writersofinstagram | writersofinstagram | author |
| | instabooks | instabooks | instabooks | life |

| | all | pos | neu | neg |
|---|---|---|---|---|
| hashtags not in all groups (20) | instagood | instagood | readersofinstagram | horror(9th) |
| | booklove | booklove | photography | nerd(26th) |
| | photooftheday | photooftheday | coffee | readersofinstagram |
| | story | story | currentlyreading | fiction |
| | readinglist | readinglist | booklove | stephenking(31st) |
| | kindle | stories | bookstagramfeature | currentlyreading |
| | stories | kindle | bookblogger | history(34th) |
| | photography | bestoftheday | write(37th) | photography |
| | readersofinstagram | pages | fiction | sad(39th) |
| | inspiration | page | travel | travel |
| | goodreads | paper | quote(41st) | novel |
| | pages | imagine(39th) | bookporn | bookblogger |
| | bookstagramfeature | happy(40th) | readers | harrypotter |
| | bookblogger | inspiration | summer(44th) | war(44th) |
| | coffee | goodreads | amreading | amreading |
| | bookporn | text | harrypotter | bookstagramfeature |
| | bestoftheday | literate(46th) | readingtime(47th) | coffee |
| | page | nook(48th) | novel | bookporn |
| | paper | plot(49th) | inspiration | photo(49th) |
| | text | booklovers(50th) | bookishfeatures(50th) | readers |

| |
|---|
| all+pos |
| all+pos+neu |
| all+neg+neu |
| neu+neg |

As shown in Table 15, thirty hashtags are common to the top 50 lists of all groups. Some hashtags are within the top 50 in two polarities, e.g., #booklove and #inspiration in positive and neutral polarities, and #readersofinstagram, #photography, #coffee, #bookstagramfeature, #currentlyreading, #fiction, #travel, #readers, #bookporn, #amreading, #harrypotter, and #novel in neutral and negative polarities. There is no hashtag overlap between positive and negative polarities, which indirectly confirms the authentication of the opinion polarity classification method. There are also hashtags only appearing in one polarity, e.g., #imagine, #happy, #literate, #nook, #plot, and #booklovers in positive polarity; #write, #quote, #summer, and #readingtime in neutral polarity; and #horror, #nerd, #stephenking, #sad, and #photo in negative polarity. The distribution of frequent hashtags in each opinion polarity reflects that certain hashtags without obvious sentimental semantic orientation are more used in a certain polarity. However, merely considering the frequency of hashtags cannot clearly capture the prominent hashtags in each group, since thirty hashtags are shared by the top 50 hashtags of each polarity. Therefore, the percentage of hashtags in each group was calculated to deepen the analysis.

The percentage of hashtags in each group (per_hg) is calculated by Equation 2:

$$Equation\ 2: per\_hg = \frac{the\ number\ of\ a\ hashtag\ in\ each\ polarity}{the\ number\ of\ this\ hashtag\ in\ all\ captions}$$

***(In this study, a hashtag is considered to be only used once in a post.)***

The percentage of positive, negative, and neutral posts is 50.5%, 5%, and 44.5% respectively. The per_hg of each of the top 50 hashtags in these groups is compared with the corresponding percentage in each group. Fisher's exact test (Bower, 2003) was employed for the comparison. The results of Fisher's exact test in Table 16 demonstrate whether a hashtag is significantly more common in a certain opinion group.

**Table 16: Results of Fisher's exact test in each polarity group**

| | hg | hg_amount | hg_per | p_value |
|---|---|---|---|---|
| **Pos** | bestoftheday** | 61190 | 98.0% | 0.000 |
| | literate** | 49235 | 95.8% | 0.000 |
| | *plot ** | *46861* | *95.5%* | *0.000* |
| | *imagine ** | *53908* | *94.3%* | *0.000* |
| | happy** | 52262 | 94.1% | 0.000 |
| | *nook ** | *47148* | *93.9%* | *0.000* |
| | *page ** | *57141* | *92.5%* | *0.000* |
| | *paper ** | *55631* | *90.5%* | *0.000* |
| | instagood** | 140159 | 89.5% | 0.000 |
| | *pages ** | *59825* | *89.7%* | *0.000* |
| | **love**** | **267743** | **89.0%** | **0.000** |
| | *stories ** | *63860* | *84.4%* | *0.000* |
| | *photooftheday ** | *87175* | *83.4%* | *0.000* |
| | *text ** | *49691* | *82.0%* | *0.000* |
| | *story ** | *76067* | *81.4%* | *0.000* |
| | *readinglist ** | *68486* | *80.8%* | *0.000* |
| | *kindle ** | *63105* | *79.9%* | *0.000* |
| | booklove** | 102068 | 78.0% | 0.000 |
| | ***author**** | ***97414*** | ***77.0%*** | ***0.000*** |
| | goodreads** | 51232 | 76.0% | 0.000 |
| | ***words**** | ***97893*** | ***76.0%*** | ***0.000*** |
| | inspiration** | 51846 | 73.1% | 0.001 |
| | booklovers** | 43434 | 72.3% | 0.002 |
| | ***literature**** | ***109867*** | ***71.8%*** | ***0.003*** |
| | ***life**** | ***57996*** | ***69.8%*** | ***0.009*** |
| **Neu** | **love**** | **29671** | **9.9%** | **0.000** |
| | **author**** | **26243** | **20.7%** | **0.000** |
| | booklove** | 27220 | 20.8% | 0.000 |
| | **words**** | **28082** | **21.8%** | **0.001** |
| | inspiration* | 18378 | 25.9% | 0.007 |
| | **literature*** | **39332** | **25.7%** | **0.007** |
| **Neg** | war** | 2259 | 58.4% | 0.000 |
| | horror** | 7245 | 45.9% | 0.000 |
| | sad** | 2498 | 34.8% | 0.000 |

***** significant level of 0.05, * significant level of 0.1***

The words marked in red in Table 16 are the top 50 hashtags shared by all opinion polarities. However, according to the result of Fisher's exact test, #love, #author, #words, #literature, and #life tend to be used in positive captions, taking into account the rate of each polarity in the overall dataset. #love, #author, #words, and #literature are significantly used in neutral

captions as well. That is to say, such hashtags are far more likely to be used in positive or neutral captions, even though they are among the top 50 hashtags in each opinion polarity. Furthermore, fifteen out of twenty-five hashtags which are overwhelmingly used in captions labeled as positive, contain no clear sentimental meaning. This finding echoes the indication from Table 15 that many hashtags with a neutral meaning are disproportionately used in positive captions. This can be attributed to personal positive attitudes towards these hashtags.

Only three hashtags are more used in negative captions: #war, #horror, and #sad. It can be concluded that current readers tend to hold negative opinions concerning these three topics on Instagram.

### 4.4.2 Result of emotion classification

The result of emotion classification is combined with opinion polarity classification. The distribution of emotion classifications is discussed under positive and negative opinion polarities, as is shown in Figure 20. Neutral polarity is excluded from emotion classification owing to the fact that these six emotions are classified into positive and negative in Table 7.

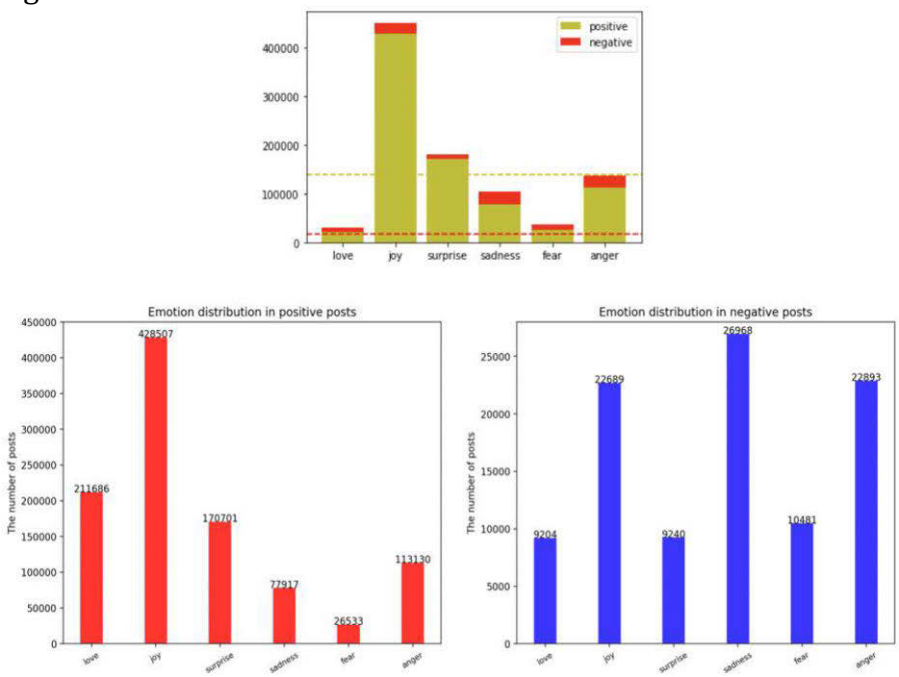



**Figure 20. Emotion distribution in negative and positive captions**

In Figure 20, the number of emotions in each group is presented. As seen from the top sub-figure, all emotions are much more common in positive polarity than that in negative polarity. In order to see the distribution of

emotions in each polarity more clearly, two sub-figures were created. In positive captions, positive emotions account for 78.8% and the emotion joy is predominant. This could imply that readers tend to feel happy, cheerful, and delighted (key feelings of joy (Shaver et al., 1987)) while publishing positive posts regarding reading on Instagram. When it comes to negative captions, negative emotions account for 68.6%. However, the number of emotion joy is larger than that of anger and fear, which means that this emotion classification model works better on positive posts.

The distribution of emotion under specific hashtags is studied. The selection of the hashtags is based on whether that hashtag is unique within a polarity or whether the hashtag is disproportionately used in that polarity. The percentage of hashtags in each emotion is calculated by Equation 3:

$$\text{Equation 3: emo}_{(\text{hg})} = \frac{\textit{the number of posts with hg in pos(neg) emotions}}{\textit{the number of posts with hg in pos(neg) posts}}$$

The result is shown in Table 17:

**Table 17: Emotion mining under specific hashtags**

| | hg | hg_amount | sad | fear | anger | surprise |
|---|---|---|---|---|---|---|
| **Neg** | horror | 7245 | 11.3% | **56.4%** | 5.7% | 10.8% |
| | nerd | 3515 | 20.9% | 8.6% | **25.3%** | 13.1% |
| | stephenking | 2992 | 15.9% | **49.6%** | 7.3% | 9.4% |
| | history | 2917 | **13.3%** | 4.1% | 8.6% | 6.1% |
| | sad | 2498 | **73.1%** | 3.6% | 13.1% | 4.1% |
| | war | 2259 | **9.0%** | 1.4% | 6.3% | 3.1% |
| | photo | 2010 | 10.7% | 3.4% | **14.9%** | 6.7% |

| | hg | hg_amount | love | joy | surprise |
|---|---|---|---|---|---|
| **Pos** | love | 267743 | 19.7% | **38.2%** | 17.2% |
| | instagood | 140159 | 13.2% | **37.3%** | 20.2% |
| | literature | 109867 | 15.2% | 22.1% | **23.6%** |
| | booklove | 102068 | **30.1%** | 28.9% | 25.2% |
| | words | 97893 | 14.5% | **25.4%** | 21.5% |
| | author | 97414 | 16.7% | 13.4% | **29.4%** |
| | photooftheday | 87175 | 10.9% | 24.0% | **24.3%** |
| | story | 76067 | 12.1% | 14.6% | **24.9%** |
| | readinglist | 68486 | 11.0% | 8.6% | **27.3%** |
| | stories | 63860 | 10.1% | 8.6% | **26.6%** |
| | kindle | 63105 | 13.8% | 11.2% | **22.0%** |
| | bestoftheday | 61190 | 8.7% | 9.5% | **27.6%** |
| | pages | 59825 | 10.0% | 6.0% | **25.4%** |
| | life | 57996 | 19.7% | **47.5%** | 15.7% |
| | page | 57141 | 8.1% | 5.9% | **27.4%** |
| | paper | 55631 | 8.1% | 7.5% | **23.7%** |
| | imagine | 53908 | 9.1% | 4.3% | **25.6%** |
| | happy | 52262 | 16.1% | **64.5%** | 13.3% |
| | inspiration | 51846 | 27.7% | **37.8%** | 9.8% |
| | goodreads | 51232 | **36.8%** | 30.3% | 22.9% |
| | text | 49691 | 6.9% | 6.4% | **26.3%** |
| | literate | 49235 | 7.6% | 3.1% | **27.0%** |
| | nook | 47148 | 8.1% | 4.0% | **22.5%** |
| | plot | 46861 | 8.1% | 2.5% | **21.2%** |
| | booklovers | 43434 | 25.5% | **30.4%** | 26.5% |

For negative posts, the emotion of sadness is outstanding for #war, #history, and #sad, while #stephenking and #horror fall under the emotion of fear the most. This is a reasonable finding because Stephen King is a

popular horror novel writer. Readers might easily get scared while reading his books. For #nerd and #photo, the emotion of anger is dominant. As for positive posts, the emotion joy is dominant in #love, #instagood, #words, #life, #happy, #inspiration, and #booklovers; #booklove and #goodreads contain the emotion of love the most. All the other listed hashtags in positive polarity have surprise as the outstanding emotion.

# 5. Discussion

This dissertation work explores the comprehension and applicability of Big Data in the context of public libraries and identifies the roles of public libraries in the context of Big Data. After demonstrating a theoretical framework for the value creation of Big Data analytics, Saggi & Jain (2018) argue that potentials should be exploited for using Big Data to create a smart city, in which public libraries could play an important role. Furthermore, Sun (2019, p.1083) advocates that the "library is in the position of knowledge service center in the knowledge society" before the technology or science is widely applied. That is to say, the library is in a good position to understand and utilize Big Data in the era of Big Data (Hoy, 2014; Huwe, 2014; Noh, 2015; Reinhalter & Wittmann, 2014). These studies correspond to the online survey results in Study 1 during which librarians made statements such as the following:

*"Libraries should be at the front and center in developing the use of Big Data"*

*"Libraries have a big and important role in dealing with Big Data"*

*"The library sits on Big Data so the subject is unavoidable"*

Sun (2019) states that the library should function as a data center rather than a document center in the context of Big Data. As was raised in the interviews with library directors in Study 1, millions of pieces of data concerning collections, users, and economics are recorded in the public library. Moreover, the good position of libraries to use Big Data is also promoted by the process of digitalization in libraries. The discipline of LIS "focuses on the component of information chain" (Robinson, 2009, p.587), which is to say that the discipline concerns "the whole communication chain of recorded information: from creation, to use, through organization, management, and dissemination" (Koltay, 2016, p.781). According to Mapulanga (2013), digitalization facilitates information creation, access, collection, development, and so on. This could lead to an increase in data volume, variety, and velocity, all of which are relevant to the advent of Big Data.

Ylipulli & Luusua (2019) state that the abundance of Big Data boosts the establishment of smart cities, where public libraries play important roles in bringing the new technologies to citizens' lives. It is therefore inevitable for Finnish public libraries as the pioneer in the field (Vakkari & Serola, 2012; Ylipulli & Luusua, 2019) to understand, manage, and utilize Big Data. Nevertheless, such an inevitable trend cannot weaken the challenge of manipulating Big Data. According to the survey and interview in Study 1, the lack of a system to handle vast amounts of data and the lack of relevant

resources are two major issues. Based on the study by Surbakti, Wang, Indulska, & Sadiq (2020), the effective use of Big Data is influenced by perceived organizational benefits and aspects, process management, data privacy and quality, security and governance, people aspects, systems, tools, and techniques. Therefore, knowing that public libraries and Big Data are a good fit is far from enough. A systematic perspective is required to strengthen the fit between Big Data and public libraries. In such circumstances, identifying roles in Big Data for public libraries, defining Big Data specifically in librarianship, and exploring use cases of Big Data application are of significance. The achievement of these three aspects fills research gaps between Big Data and public libraries and handles the methodological issue in social media studies by considering Instagram as the platform to further this doctoral dissertation. These three aspects together correspond to the overall purpose of this doctoral study: to help public libraries to realize what their responsibilities might be in the context of Big Data and to understand what Big Data is and how to apply Big Data.

According to Figure 5, the overall purpose is fulfilled by realizing three subordinate aims:

- Aim 1: To identify roles of public libraries in the context of Big Data
- Aim 2: To define Big Data specifically in librarianship
- Aim 3: To provide practical examples of Big Data applications for public libraries

Aim 1 and Aim 2 lay the theoretical foundation of the study, which helps guide Big Data applications in public libraries. Therefore, the discussion of the research findings is organized around the three subordinate aims accomplished by the four studies outlined above.

## 5.1 The roles of public libraries in the context of Big Data

Study 1 was conducted to fulfill Aim 1. Section 2.3.2 highlights that studies have noticed the benefits of applying Big Data since the year 2015. However, current studies lack focus on demonstrating the roles and responsibilities of public libraries to meet Big Data. Nine roles put forward by this study can fill such a research gap. As is reviewed in Section 2.1, Finnish public libraries are highly developed and good examples to public libraries in other countries. Since Study 1 was designed and approached in Finnish public libraries, therefore the outcome of this study could contribute to the development of public libraries in the era of Big Data in other countries.

To date, there have been no studies systematically identifying roles regarding public libraries for the advent of Big Data. Thus, Study 1 was designed with an online survey to test the opinions on roles attained from the existing literature and experienced librarians. In order to ensure the quality of these identified roles, eleven semi-structured interviews for

library directors and managers were conducted as well. There are eight roles tested in the survey. The result of the interviews further rationalizes these eight roles in the survey, which consolidates the responsibility of these eight roles for public libraries. One more role – facilitator – was put forward by the interviewees based on the reality of Finnish public libraries.

The hierarchy of these nine roles is displayed in Figure 21:

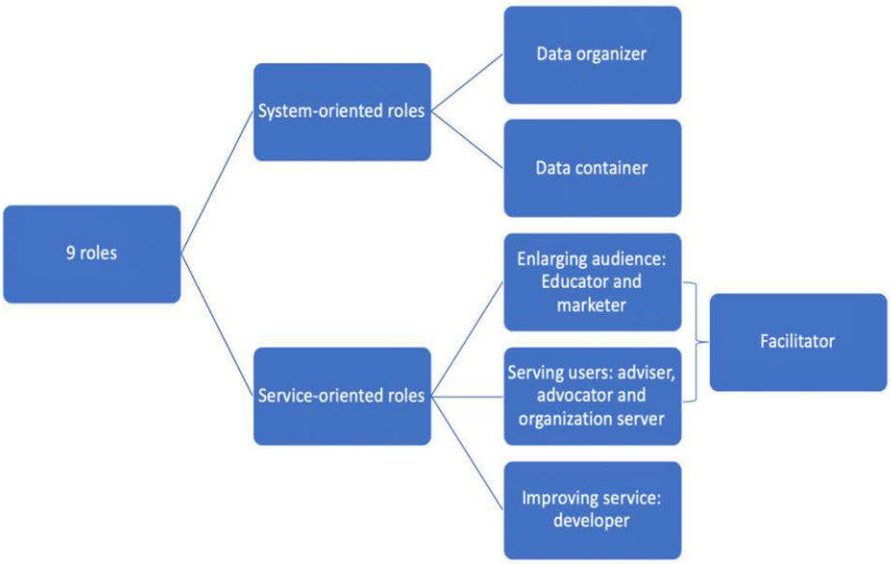



**Figure 21. The hierarchy of nine roles**

System-oriented roles (data organizer maintaining Big Data sets and data container storing Big Data) concern to practices of public libraries in managing Big Data, which emphasize the sustainable development of data, archiving, and storage of data for future use. Service-oriented roles are classified into three groups based on their functions: (a) educator and marketer to enlarge the audience, (b) adviser, advocator, and organization server to serve library patrons in using Big Data, and (c) developer to enhance services with the application of Big Data. There is a final role, facilitator, which is advocated when the library has very limited resources to realize the roles of serving users and improving services.

### 5.1.1 Theoretical implications of identifying roles of public libraries in the context of Big Data

Survey respondents hold conservative opinions towards system-oriented roles. However, library directors and managers are positive about these two roles. According to interviewees, system-oriented roles should be mainly undertaken by large public libraries, which partially explains the unclear attitudes towards them in the survey. Respondents from small libraries

hardly see the necessity of these two roles in their own libraries, thus providing a lower score regarding system-oriented roles. During the interviews, the National Library of Finland was mentioned as the exclusive library to store data for all public libraries in Finland. Library directors and managers interviewed similarly recommended that the main public libraries in each municipality area should organize and maintain data for the secondary libraries. It can be concluded that the scale of a public library needs consideration before deciding which roles that library should undertake in the context of Big Data.

Compared with system-oriented roles, service-oriented roles receive stronger agreement from the survey respondents. Librarians slightly agree with the roles of educator, marketer, advocator, and adviser in the survey, owing to their claim that library users could benefit from Big Data. In the answers to the open question of the survey, words such as "help individuals" and "guidance" are mentioned several times, which indicates that librarians realize their responsibility to introduce Big Data to citizens. These four roles are also highly agreed upon by the library directors and managers during the interview. The roles of organizer server and developer are also recommended by the interviewees because they foresee the bright future of public libraries having these two roles in the generation of Big Data to enhance their services and boost connections with other organizations. To sum up, librarians, library directors, and managers are positive towards service-oriented roles because they realize their duty to serve people and the benefits of Big Data utilization. As is presented in Figure 21, the service-oriented roles are classified into three groups:

1) To enlarge the audience of Big Data. The point of the educator role is to help citizens understand what Big Data is. This role can be reflected by studies (Hoy, 2014; Jaeger et al., 2012; Ylipulli & Luusua, 2019) advocating that public libraries provide training regarding new technology to ensure citizens' access and the utilization of such technology. As is stated by Hoy (2014, p. 324), "librarians will need to help their patrons understand what Big Data can and cannot do, and how it can best be used to achieve their research goals." The meaning of marketer is to broadcast the concept and benefits of Big Data, which is also encouraged by Gordon-Murnane (2012) and Wittmann and Reinhalter (2014). According to the interview, public libraries should promote the opportunities for our lives brought by Big Data. The agreement on these two roles indicates that public libraries ought to function as a bridge for citizens to the context of Big Data.
2) To help users to utilize Big Data. Summarized from the interview, users of public libraries are individuals, communities, and organizations. Three roles (adviser,

advocator, and organization server) are identified for public libraries to assist the utilization of Big Data by users. These three roles echo the study by Hoy (2014) that calls for libraries to be a technology adaptor in the context of Big Data to guide users to access valuable databases, manage their own big datasets, and so forth. These three roles could be considered the next step after realizing the library's role as educator and marketer (explaining to users what Big Data is). The adviser, advocator, and organization server teach library users what Big Data can do.

3) To improve library services. All service-oriented roles center on providing services for library users. The roles in the aforementioned two groups emphasize the comprehension and application of Big Data for users. The developer role is designed for public libraries to employ Big Data for their own services. As is mentioned in Section 2.3.2, Big Data has been applied in the public library to enhance library strategy and decision making, which manifests that the role as developer has been realized by public libraries in the context of Big Data. According to the survey and interview, most librarians, library directors, and managers are keen to see practical examples of Big Data application, under which public libraries undertake the role of developer. Such interest encouraged this doctoral project to provide two cases of Big Data application.

### 5.1.2 Practical implications of identifying roles of public libraries in the context of Big Data

One important practical implication of identifying public library roles in the context of Big Data is that it will be essential to prepare corresponding resources for libraries. Realizing the necessity of system-oriented roles requires public libraries to arrange devices and techniques to store and maintain a large amount of data. Service-oriented roles require not only techniques but also platforms to communicate with library users concerning Big Data. However, the reality of resource limitation has been a challenge for public libraries. Therefore, the combined role of a facilitator was put forward by interviewees. The facilitator role integrates the functions of the four roles of educator, marketer, adviser, and advocator and positions public libraries as a platform where external professionals can be introduced to citizens. Compared with these four separate roles, the concept of a facilitator is more consistent with the current situation in public libraries. As the available resources are too limited to accomplish all four roles, it would be wise to let other people help libraries accomplish this task. In addition, no matter which role(s) a public library undertakes, legislation issues such as copyright and

the right to use personal data should be clarified in advance. Otherwise, the willingness of libraries to use Big Data in practice will be diminished.

## 5.2 The definition of Big Data in librarianship

Study 2 was conducted to reach Aim 2: To define Big Data specifically in librarianship. According to Section 2.2, current studies tend to generally define Big Data considering the characteristics of Big Data. However, there is a gap in research to comprehend Big Data for a specific field. Therefore, a definition of Big Data for librarianship achieved in this doctoral study could be considered a pioneer example to mainly focus on defining Big Data in a specific field. In addition, such a definition could be helpful for public libraries when undertaking the roles of educator and marketer to bring Big Data to citizens' lives. A total of 35 definitions of Big Data were collected from 124 articles published in 179 English-language journals and eight conference proceedings in the field of library and information science (LIS). Key aspects, word frequency, and the meaning similarity between pairs of definitions are highlighted. A summarized definition of Big Data for librarianship is achieved.

### 5.2.1 Theoretical implications of defining Big Data in librarianship

By analyzing collected definitions, it was possible to identify two classifications of Big Data definitions in LIS literature: data-oriented and ability-oriented. Data-oriented definitions emphasize that Big Data is data or information with corresponding features (e.g., increasing fast). Ability-oriented definitions consider Big Data to be the techniques to manage large amounts of data. In the word frequency list in Table 10, data and information are mentioned the most in the collected definitions, because some definitions do not specify the difference between data and information, which enlarges the domain to define Big Data in librarianship. Definitions in both classifications consider data the nexus: Big Data is either data itself or the ability to process data. V's that are used to characterize Big Data in previous studies can be easily found in the collected definitions and are reflected in definition content analysis and word frequency. This indicates that Big Data in librarianship shares the same characteristics as mentioned in previous studies. Four V's (volume, velocity, variety, and veracity) are clearly stated in some of the analysed definitions. The advantages of Big Data are reflected in eight definitions, which highlight another V of Big Data: Value (Wamba et al., 2015). Therefore, it is concluded that in librarianship, Big Data has five characteristics: volume, velocity, variety, veracity, and value.

The demand and challenge aspects identified in the content analysis of Big Data definitions reveal that there are difficulties for current public libraries to overcome as they step into the era of Big Data. These two aspects are relevant to the study by De Mauro et al. (2016), which states that technology to use Big Data and techniques to process it are two themes of Big Data. It is

necessary to involve the technology perspective when defining Big Data for librarianship. Studies outline technological challenges and requirements for libraries to apply Big Data:

- Lack of professionals (Gordon-Murnane, 2012), which was also raised during the interview and online survey of Study 1.
- Lack of supportive infrastructure (Hoy, 2014), such as systems for data storing (Fuller, 2015).
- Lack of a formal data modeling process, which is relevant to managing, accessing, and describing Big Data.

Furthermore, personal data is (or could be) recorded in the public library. The problem of protecting user privacy while utilizing user-related data is challenging. Since there is no practical standard to decide whether a dataset falls into the scope of Big Data (e.g., data, where the volume is greater than a certain amount, or is growing at a speed higher than a certain rate), it is an optional standard to concern the challenge and the demand such a dataset brings, (e.g., if a dataset causes challenges for a public library to manage, should that be defined as Big Data?). By conquering the challenge and meeting the demand, public libraries could develop technical solutions. This implies a potential advantage of Big Data for public libraries: bringing development.

To summarize, the following Big Data definition for librarianship can be proposed: Big Data refers to data with a large size, rapidly growing speed, and various types, which can complicate data handling techniques but also boost the creation of technological solutions. Value is generated by the proper operation and use of Big Data.

### 5.2.2 Practical implications of defining Big Data in librarianship

According to the definition of Big Data generated by this study, certain techniques and operations are needed to use Big Data. Therefore, the skills of librarians should be developed so that librarians can perform well in the context of Big Data. The practical implication of defining Big Data in librarianship echoes certain roles identified by Study 1, which further demonstrates the feasibility of those suggested roles.

The advent of Big Data makes it inevitable that librarians in the public library will need to handle datasets that hold a great amount of data, increase rapidly in size, and contain various types of data. In the meantime, librarians need to assure the quality of data saved in datasets so as to facilitate the use of the data in the future. Simply put, librarians need to confront data with the four V characteristics of volume, velocity, variety, and veracity. With the growth of digitalization in the public library, the amount of data saved in a library's database, such as borrowing histories, collection information, and user profiles, is rapidly exceeding the database's capability. Moreover, the widespread utilization of social media and the increasing number of mobile devices accessing the library aggregates the challenge of

data volume. Along with various ways to generate data, the varieties of data in the library system would also be enriched.

In order to overcome such challenges, cloud computing, which refers to a location where data is stored in a storage mechanism not located in the local area network (Affelt, 2015, p.41), is recommended for current public libraries, owing to a trend of outsourcing data to the cloud (Wang et al., 2010). According to Wang et al. (2010), cloud computing relieves the burden of storage management, provides easier access, and saves on hardware, software, and personnel maintenance costs. Therefore, the skills to employ cloud computing for library data storage are relevant in the context of Big Data. Due to the continuous generation of data, skills for efficiently collecting data should be gained by librarians, especially those who work in municipal public libraries, because these libraries should undertake the role of data container and data organizer for their subordinate libraries (Zhan & Widén, 2018). In addition, skills for storing different types of data (for example text, image, video, and statistics) and skills for cleansing the noise from big databases are also required for librarians providing technical services. With such help, the successful application of Big Data within a library can be assured.

According to Study 1, public libraries should work as a developer to improve their services with the application of Big Data. Therefore, skills for analysing a large amount of data are required for librarians as well. Ahmad et al. (2019) recommend that librarians develop their skills (e.g., data analysis, data visualization, programming language, and coding) so as to cope with Big Data. For example, librarians need to help individuals find and use library materials, understand users, and provide the most relevant resources. Knowledge of programming languages such as Python or R to analyze Big Data would be useful in order to undertake user service duties. Additionally, since knowledge management is required by librarians, skills for maintaining and utilizing knowledge generated from Big Data analytics are also needed. For some librarians, their skills in handling the relevant issues of Big Data are salient, including the privacy issue of cloud computing, the security issues of Big Data protection, decisions on which part of Big Data could be used for value creation, and governance issues regarding the whole process of Big Data application. The duty of librarians undertaking such administrative services could be characterized as harnessing Big Data at the management level (Nasser & Tariq, 2015; Wang et al., 2010).

## 5.3 The application of Big Data on Instagram for public libraries

Aim 3 focuses on providing practical examples of Big Data applications for public libraries. Study 3 and Study 4 were conducted to fulfill this aim. In this doctoral study, Instagram is selected as the platform to collect data. As is pinpointed in Section 2.5, most studies concerning Instagram utilization in

LIS were conducted in the context of academic libraries. Therefore, Study 3 and Study 4 will enrich the using environment of Instagram since both of them emphasize public libraries as the subject to employ Instagram. According to Section 2.5, the use of Instagram in librarianship can be classified as user engagement, library showcasing, keeping content dynamic, and fundraising. This doctoral study chooses a new angle to discuss the use of Instagram: hashtag. Considering the research gaps identified in Section 2.5 and 2.6, the findings in the study can; a) enlarge the domain of Instagram use in public libraries from a hashtag function point of view; b) demonstrate a new way to use hashtags: hashtag organization; and c) enrich hashtag research by analyzing hashtags from Instagram.

The fulfillment of Aim 3 can broaden the scope of library research in the field of social media and Big Data. Users of social media can be both library customers and content producers (Rasmussen, 2016). When explaining the concept of the participatory library, Cuong Nguyen, Partridge, & Edwards (2012) posited that library users could take on various roles and tasks, such as organizing information resources by tagging, rating, and bookmarking and serving other users by answering questions and recommending materials. For libraries, users' organization of information resources and provision of services to other users represent user-generated content. Thus, captions and comments created by library users on Instagram could be valuable resources for libraries to advocate user participation. The findings of Study 3 and Study 4 assist to understand current readers and facilitate communication between users and public libraries, which are beneficial to the creation of participatory libraries.

According to Figure 5, the realization of Aim 1 and Aim 2 provides guidance for the fulfillment of Aim 3. Study 1 advocates that developer is one of the most important roles for public libraries to undertake in the context of Big Data. This role emphasizes the practical utilization of Big Data for the benefits of public libraries from a service improvement point of view. Moreover, the definition of Big Data for librarianship outlines that value generation ought to be a characteristic of Big Data in librarianship. These two viewpoints together suggest that the application of Big Data could start with exploring possibilities to enhance library services.

Based on the result of Study 3 and Study 4, it can be concluded that by analyzing hashtags in Instagram captions, public libraries can arrange their resources, improve and encourage reading activities, and engage with users. During this process, new services can be developed, which realizes the library's role of developer in the context of Big Data and strengthens the definition of Big Data specifically for librarianship. That is to say, the application of Big Data can also help to clarify the roles and responsibilities of public libraries in the context of Big Data and reinforce the comprehension of the Big Data definition for librarianship.

### 5.3.1 Hashtag analysis to enhance the communication between public libraries and users

The communication function of hashtags is consolidated in this study through the effective organization in the caption. Nearly seven million library-related captions were collected within the space of four and half months. Practical insights about hashtag use on Instagram are imparted so as to outline an efficient pattern for public libraries to communicate with their users.

After cleaning the dataset, 2,561,424 captions were used for further analysis. Considering the distribution of hashtags in a caption, 15 hashtag locations, which denote the location of all hashtags in a caption, were identified. Based on the statistical description of each variable, two types of regression models were established to demonstrate which hashtag location would achieve the most "likes" and "comments". Possible ways to organize hashtags are provided by this study to attain more "likes" and "comments" from library users on Instagram, demonstrating how communication between the public library and users on Instagram could be enhanced via wise placement of hashtags in the caption.

According to Table 11, the average number of hashtags used in library-related posts on Instagram are 16, which reflects a habit among library users to use more than one hashtag for their posts. This average number also can reflect the tendency of Instagram users to put more than one hashtag in their posts. Both insights pinpoint the significance of hashtag organization on Instagram.

Based on the comparison of six regression models (Tables 13 and 14), it can be concluded that hashtag locations play a more important role in attaining "likes" and comments than time difference (TD) or the hashtag amount (HC) and the word amount (WC) of a caption. It should be highlighted that TD is only significant in getting more "likes" when there is more than one hashtag in the caption, and TD has no influence in obtaining more comments. Although TD is significant in certain circumstances, the coefficient value is rather small. Therefore, it is not the case that the longer the post exists on Instagram, the more "likes" and comments the post will attain. In addition, the number of comments has a positive influence on the number of "likes". However, receiving more "likes" will not lead to an increase in comments. As such, from the perspective of strategy, organizing hashtags to get more comments should be prioritized.

When there is only one hashtag in the caption, it is recommended to avoid using it in the front, which has no positive relation to attaining either "likes" or comments. For boosting comments, a single hashtag should be used in the middle of the caption. When it comes to increasing "likes", it should be put at the end.

When the number of hashtags accounts for more than 80 percent of a caption, it is wise to create the caption with a few non-hashtag words rather

than only hashtags. In this way, the post can attain more "likes" and comments.

More than 75% of captions have more than one hashtag but a hashtag percent of less than 80%, which includes ten types of locations. As reflected by Figure 17, more than 60% of captions have these hashtags all placed consecutively at the end. The location with the second-most captions is ME (more at the end), which accounts for 12.75%. ME, AM (all in the middle), AE (all at the end), and LF (less in the front) are the locations that lead to both more comments and more "likes". Considering the value of the coefficient of these four locations, ME is the best. There is another location, S not F (scattered but not in the front), which is only good for getting comments. BCE (beginning and centralized at the end) is merely good for getting "likes". According to the result of Fisher's exact test on captions with zero comments, AB (all at the beginning) is the only location resulting in a significantly greater number of captions without any comment.

Based on these practical considerations, three suggestions can be provided:

- Irrespective of the number of hashtags in the caption, one should avoid using them mainly at the beginning of the caption.
- It is not wise to create a caption with all hashtags.
- Increasing the number of comments should be prioritized, and any hashtag(s) should be mainly used in the middle of the caption.

### 5.3.2 Hashtag analysis to understand readers

In this study, ad-hoc sentiment analysis was conducted on millions of captions with the hashtags #read and #reading. The opinion polarity of each caption was identified. Since this study intends to identify topics that current readers either like or dislike, captions classified as neutral opinion polarity are not further discussed. Consideration of the discussion will concentrate on hashtags more used in positive and negative captions. The emotions within these two polarities were also highlighted in order to provide a thorough analysis of current readers' sentiments towards various topics.

After sorting the captions into groups based on the opinion polarity and emotion classification, prominent hashtags in that group were outlined so as to represent the main topic. Owing to the function of hashtags to denote the content of the caption, this study concentrates on the frequency of hashtags in each sentiment group, rather than review the whole content of each caption. According to Table 6 and Figure 14, the models to identify opinion polarity and classify emotions reach reasonable accuracy, which consolidates the findings of this study.

As illustrated in Figure 20, the emotion joy is most common in positive captions, and the total number of captions with this emotion is likewise much higher than captions with one of the other five emotions. As is explained in Table 7, positive outcomes are the main cause of the emotion

joy, and it can be inferred that current readers tend to achieve positive outcomes during reading easily. Therefore, public libraries should provide services to help current readers to make a record of these positive outcomes in order to enhance this tendency of joyful reading. For instance, an electronic record of books that have been read by a library user can be provided when the user logs into the library system. Such a record can be considered the presentation of the user's reading achievement. Such an achievement could be helpful to keep their emotion of joy in reading and thus promote the acceptance of libraries' reading-related services.

Table 16 lists 25 hashtags, attained from Fisher's exact test, that frequently appear in positive captions. After considering the textual meaning of these hashtags, they fall into four categories, which can be summarized as follows:

- Reading materials: #literature, #words, #author, #story, #readinglist, #stories, #pages, #page, #text, and #plot.
- Feelings and reflections during reading: #love, #booklove, #imagine, #happy, #inspiration, #goodreads, and #booklover.
- The location and the way to read: #kindle, #paper, and #nook.
- Life: #photooftheday, #bestoftheday, and #life.

These four categories reflect aspects towards which that current readers are positive.

First of all, when people read some materials, especially those that are surprisingly better than expected, they tend to hold positive opinions. Such a conclusion can be indicated by the finding that for all the hashtags in the category of reading material except #words, the dominant emotion is surprise. Dewan (2013) advocates that academic libraries should shift their attention from the quality of reading materials to the quality of reader experiences. However, such a shift should not be advocated in public libraries, because current readers like materials with surprisingly good quality, which boosts their positive posts towards reading, thus indirectly improving their impression of the library.

Another group of topics towards which readers hold positive opinions is their feelings and emotions generated during reading. These feelings and emotions could be attributed to two basic emotions: love and joy. Considering the antecedent of these two emotions, readers would feel positive when they establish a fine interaction with the text or achieve positive outcomes during reading. Therefore, public libraries should help readers relate their personal needs to the content of the book so as to achieve readers' own goals. In this way, public libraries can improve users' reading experiences. Meanwhile, it is also necessary to implement better solutions to know library users.

Topics regarding where and how readers accomplish their reading activities are popular among positive captions. #nook denotes a peaceful

environment for reading. #kindle and #paper are different platforms for reading (#kindle means digital reading device and #paper means traditional ways to read). The significantly high frequency of all three hashtags in positive captions suggests that current readers pay extra attention to the physical reading environment and the reading mode. The high frequency of #nook echoes the studies by Durant & Horava (2015) and Waxman, Clemons, Banning, & McKelfresh (2007), both of which emphasize that libraries should be a comfortable and restorative place rather than a place for community, socialization, and intellectual discussion. According to the study by Durant & Horava (2015), two factors lay the foundation of the reading trends in the future: easy access to information and the reading preference changing from printed books to ebooks. Both factors could explain the popularity of #kindle in positive captions. The main emotion relating to #kindle is surprise, which implies that the experience of using this new reading mode exceeds readers' expectations. However, the prominence of #paper in positive captions also highlights that the emphasis on digital reading materials should not be accompanied by diminishing the number of printed books. Therefore, the findings of this doctoral study do not recommend public libraries decrease the number of their printed collections.

The use of #photooftheday, #bestoftheday, and #life in positive captions indicates that people tend to generate positive sentiments when they realize the beauty of life via reading. This reflects that current readers would like to use reading as a way to enjoy their life. Therefore, public libraries should help users to achieve such a goal.

After summarizing these positive captions and their corresponding major emotions, topics that current readers like are: (a) reading materials that surpass readers' expectations, or materials with which readers can establish a personal bond, and materials through which positive outcomes are achieved; (b) pleasant reading environments or habits; and exclamations of life realized through their reading. Equipped with the knowledge of these topics, public libraries could better prepare their collections, arrange physical environments, provide suitable resources for various reading preferences, and encourage citizens to read.

There are fewer negative captions than positive ones, and the same is true of prominent hashtags in negative captions. Only three hashtags (#war, #horror, #sad) are disproportionately more used in negative captions. It can be concluded that compared with positive hashtags, people reveal less about their personal feelings in negative captions on Instagram, given that fewer hashtags related to personal feelings are used in negative captions. These negative captions mainly concern the content or genre of what people are reading.

One thing that should be highlighted is that even if these captions are labeled as negative, it does not mean that readers dislike the content or genre. For example, most captions with #stephenking are classified as

negative. However, it cannot be said that current readers hold negative opinions towards this author. It merely indicates that his books are horror stories and that words used to describe evil or horrific things tend to be classified as negative based on the dictionaries used in VADER and TextBlob. Therefore, it comes as no surprise that captions with #sad and #horror hashtags are disproportionately classified as negative.

#war is mainly used in negative captions with the main emotion sadness. The unpleasant outcomes of war make people feel sad when they read relevant materials. Thus, it can be concluded that current readers do feel a dislike of topics connected with wars. Such negative sentiment about wars should be outlined by public libraries in order to let citizens know and understand the undesirable outcomes of wars.

### 5.3.3 Theoretical implications of hashtag analysis

This dissertation work enriches research on employing hashtags pragmatically on social media. Current studies mainly focus on the best number of hashtags to achieve interaction (Cooper, 2016; Lee, 2016), the use of retagging to add new meanings to hashtags in order to broaden the community (Oh, Lee, Kim, Park, & Suh, 2016), and ways to increase the contextual meaning of hashtags to achieve marketing goals (Bunskoek, 2014; Harkai, 2018). This study provides a new angle: hashtag organization.

The wise usage of hashtags could increase the visibility of social events (Wang, Liu, & Gao, 2016). According to the study by Wang et al. (2016), certain co-occurrences of hashtags are more popular among event participants. Therefore, they conclude that the spreading of social events on social media should be considering the characteristics of hashtags, thus increasing information virality, which is conceptualized as "the capacity of individuals and organizations to share information and successfully mobilize collective attention, as well as the ability for messages to connect diverse networks" (Wang et al., 2016, p.851). The result of this doctoral study puts forward another way to enhance information virality: hashtag organization. That is to say, certain ways of organizing hashtags could gather more "likes" and "comments", both of which reflect better information sharing and collective attention. In light of the definition of information virality, it can be claimed that hashtag organization could ensure information virality as well.

When discussing the revolution of library models, Noh (2015) states that in the future, libraries will be "intelligent libraries" and "massive data libraries". The arrangement of the presented study provides insight on how to build an intelligent library and massive data library. Additionally, apart from the role as developer, the hashtag analysis to enhance communication between public libraries and users also sheds light on another role advised by Study 1, which is adviser. The findings of hashtag organization study could advise citizens and organizations how to achieve more "likes" and

comments for their own business with the help of Big Data analytics, which corresponds the role as an adviser for public libraries in the context of Big Data (i.e., to offer advice for problem-solving from the perspective of Big Data).

The hashtag analysis to outline topics that current readers like or dislike reinforces the point summarized in Section 4.4, which is that hashtags can reveal personal feelings and emotions. For instance, #booklove is more used in positive captions, and most of the captions reveal the emotion love, while #horror is more used in negative captions, and over half of the captions denote fear as the main emotion. The opinion polarity and emotion of both hashtags can also be reflected by the contextual meaning, which also implies the feasibility of the applied machine learning algorithms.

According to Figure 20, the number of positive captions is much larger than that of negative captions. Since textual captions can represent the main content of the image or video (Zhong, Zhang, & Jain, 2000), it can be concluded that current readers would mainly like to generate positive content on Instagram to express their reading activities. In 2017, the study by Shcherbina (2017) endorses the view that people are encouraged to read and express themselves by the development of Web 2.0. The existence of millions of captions containing #reading and #read on Instagram could be attributed to the fact that people are encouraged to express themselves while reading. This encouraging tendency can be said to have contributed to the largely positive content about reading that can be found on Instagram. Therefore, the finding of Study 4 implies that public libraries could develop more events for citizens to convey their opinions and feelings towards a book so as to promote reading activities.

### 5.3.4 Practical implications of hashtag analysis

The studies included in this dissertation regarding Big Data applications contribute to good practices for libraries. Studies (Abidin et al., 2013; Cahill, 2011; Carlsson, 2012; Young, 2016; Young & Rossmann, 2015) have demonstrated that social media bring great possibilities for libraries to establish communities through communication via social media (Ganster & Schumacher, 2009) and information sharing on social media (Phillips, 2011; Young, 2016). Suggestions made in this study for attracting more user attention (indicated by the number of "likes") and encouraging user communication (indicated by the number of comments) are useful for library community-building activities. If a public library makes a post on Instagram, an effective hashtag organization could help the post could obtain more "likes", which in turn implies that more citizens would see this post, thus ensuring information sharing. An effective organization could also lead to more replies from users, which increases the opportunities of the library to communicate with people. This increase in visibility and engagement facilitates community establishment.

Realization of the application to understand readers could be considered one practical way to strengthen the public library's role as facilitator as identified in Study 1. The goal of the facilitator is to introduce external professionals to library users. The method used to outline current readers' likes or dislikes could also be applied by public libraries to understand what users really would like to know and thus to make sure that they can bring suitable external professionals to interested users.

# 6. Conclusions and Limitations

The doctoral project focuses on helping public libraries realize what their responsibilities might be in the context of Big Data and gain a better understanding of what Big Data is and how to apply Big Data. As a rising research area, few studies are combining Big Data and public libraries, not to mention those considering Instagram the platform to collect data to create Big Data applications for public libraries. The accomplishment of the present study fills research gaps in bringing Big Data to public libraries, enriches the content of Big Data applications and Instagram applications in public libraries, and handles the uneven spread of research in social media studies regarding the single-platform prevalence. Furthermore, the value of Big Data in this study is reflected by the analysis of hashtags in millions of library-related Instagram captions. The hashtag was chosen as the focal point of this doctoral study due to the universality of hashtag use on Instagram. Millions of captions were collected and analysed. Currently, data used by libraries for Big Data analytics can be characterized neither as very big nor fast-growing. The volume of data used in this doctoral study regarding Big Data application is much bigger than data used in earlier studies on Big Data applications in libraries. Furthermore, the number of captions is dramatically increasing on Instagram. Both of these factors make the data in this doctoral study more relevant in Big Data research than data from previous studies. The analysis was conducted in terms of basic functions of hashtags: indicating the semantic domain and communicating within a community. The result of the analysis indicates a novel way to use hashtags: hashtag organization and a creative way to know library users: sentiment analysis on hashtags.

Considering the current research status, inductive approaches were employed to fulfill the overall purpose. Moreover, both qualitative and quantitative methodologies were used considering the research question and the amount of analysed data. Three subordinate aims and four research questions were put forward to ensure the achievement of the overall purpose, which are fulfilled by four studies and cover (a) roles for public libraries in connection with the advent of Big Data, (b) comprehension of Big Data in librarianship, and (c) practical examples of Big Data application. In the end, the research purpose is achieved, and four research questions are well answered by the accomplishment of four studies.

## 6.1 Conclusions

There are nine roles of public libraries in the context of Big Data pinpointed. After conducting an online survey and interviewing 11 public library directors and managers, eight roles are identified. These eight roles are classified into two broader orientations: system-oriented roles (data organizer and data container) and service-oriented roles (educator,

marketer, adviser, advocator, organization server, and developer). System-oriented roles are dependent on library size. That is to say, only big public libraries should undertake these two roles and share their resources with their subordinate libraries. Service-oriented roles also require resources, which cannot be ensured by any public library owing to the reality of resource limitation. Therefore, another role – that of facilitator – was put forward by the library directors, which links external resources to the right people without libraries being directly responsible for providing the resource in question.

A definition of Big Data in a librarianship context was formulated in this study so as to provide a better comprehension of Big Data as used in libraries and thus fill this research gap. After analyzing 35 Big Data definitions used in LIS studies, the following definition was put forward: Big Data refers to data characterized by its large size (volume), rapid growth (velocity), and various types (variety), which can complicate data handling techniques (veracity) but also boost the creation of technological solutions. (Value) is generated by the proper operation and use of Big Data. It can be concluded from this definition that Big Data in librarianship shares the five V characteristics which are highlighted in the definition. The definition outlines the skills that librarians need in order to handle tasks relating to Big Data. They need analytic skills to attain value through a proper analysis process. They also need computing skills to handle issues regarding storing and maintaining data of various types, a rapidly increasing speed, and huge volume. The skills to protect user privacy are also required.

Two quantitative studies (Study 3 and Study 4) were conducted in order to provide practical examples for (Finnish) public libraries on the applications of Big Data.

The review in Section 2.6 identified no prior research on how to use hashtags effectively for public libraries, which motivated the conduction of this study. The communication function of hashtags is emphasized to outline effective ways to organize hashtags in library-related captions on Instagram in order to achieve more "likes" and comments. It is implied that the location of hashtags in the caption could influence the comments and "likes" that the caption attains. Pieces of advice for organizing hashtags in the caption are:

- Irrespective of the number of hashtags in the caption, one should avoid using them mainly at the beginning of the caption.
- It is not wise to create a caption with all hashtags.
- Increasing the number of comments should be prioritized, and any hashtag(s) should be mainly used in the middle of the caption

With sentiment analysis of readers' opinions and emotions in this study, prominent hashtags with dominant emotions in positive and negative

captions are identified. Summarization of the textual meaning of these hashtags can be used to identify topics that readers like or dislike.

For positive captions, 25 hashtags are used significantly more than others: #literature, #words, #author, #story, #readinglist, #stories, #pages, #page, #text, #plot, #love, #booklove, #imagine, #happy, #inspiration, #goodreads, #booklover, #kindle, #paper, #nook, #photooftheday, #bestoftheday, and #life. Based on the meaning of these hashtags, four categories of hashtags are outlined: reading materials, feelings and reflections during reading, the location and the way to read, and life. It can be concluded that current readers like reading materials that are surprisingly better than readers' expectations or that readers can establish personal bonds with or achieve positive outcomes from; fine environments or pleasant manners for reading; and exclamations of life realizing via reading. As for negative captions, there are only three outstanding hashtags: #war, #sad, and #horror. Despite the high frequency of words describing sad and horror books, it can be concluded that current readers most clearly dislike topics associated with wars.

The roles provided by Study 1 and the Big Data definition summarized in Study 2 highlight a direction to conduct these two quantitative studies: to enhance library services. Through these two quantitative studies, the findings of Study 1 and Study 2 can be re-examined and consolidated as well. The two quantitative studies strengthen the role of the library as developer, adviser, and facilitator. Moreover, these two studies can help librarians further understand the Big Data definition. All in all, such relation deepens the connection of the four studies in this doctoral research project.

## 6.2 Limitations and expectations for future studies

Although contributions have been made by this doctoral research with valid methods, there are still limitations.

This study involved surveying and interviewing librarians and public library directors only in Finland, which limits the applicability of this role study. In the future, wider audiences should be included. Furthermore, the study covers all identified roles equally. When it comes to these roles, emphasis might differ from library to library. Therefore, studies on the ranking of these identified roles, based on their importance within different public libraries, would be valuable.

The number of analysed definitions is limited in the study. Moving ahead, future studies should cover a collection of definitions within a wider domain. Meanwhile, the skills of librarians in the context of Big Data are generally discussed. Future studies should also conduct a more specific analysis regarding librarian skills for different types of libraries in the context of Big Data.

The content of the analysed captions is restricted to library-related topics in the present study. Therefore, the applicability of the study is limited. For

instance, more topics should be included to broaden the applicable scope of hashtag organization. Secondly, the association factors which can generate more comments and "likes" or affect sentiment analysis results should be considered and further discussed. The study does not differentiate between captions made by individuals and organizations, which could impact on attaining more comments and "likes", thereby should be considered in future studies. As is highlighted in Section 5.3.2, captions classified as negative are more relevant to book content rather than to readers' opinions, which is a major limitation of sentiment analysis in this study. In future studies, more filters should be added to help collect captions mainly reflecting readers' opinions. Thirdly, even though proper steps have been taken to clean the collected captions, there is still room to improve this process. As addressed by La Sorte et al. (2018), it is always a challenge to clean data. Therefore, more sophisticated methods should apply to clean data in future studies. Finally, a comparison study should be conducted to further demonstrate whether library-related captions share the same patterns of effective hashtag organization as captions on other topics.

The domain issue proved a significant challenge for Study 4, since there is no ready-to-use dataset for sentiment analysis generated on Instagram data. In addition, the emotion classification model works better on positive captions than negative captions, reflected by the findings visualized in Figure 20, where the emotion joy ranks as the third highest in negative captions. Such a result could affect the credibility of the machine learning model used in this study. In the future, a dataset for sentiment analysis on Instagram captions should be designed. Moreover, the analysed captions come from a limited time period. In future studies, a more advanced method to collect and analyze data should be developed so as to monitor changes in liked or disliked topics over time.

# References


Abbasi, A., Sarker, S., & Chiang, R. H. L. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, *17*(2), I.

Abdullah, N., Chu, S., Rajagopal, S., Tung, A., & Kwong-Man, Y. (2015). Exploring Libraries' Efforts in Inclusion and Outreach Activities Using Social Media. *Libri*, *65*(1), 34–47. https://doi.org/10.1515/libri-2014-0055

Abidin, M. I., Kiran, K., & Abrizah, A. (2013). Adoption of Public Library 2.0: Librarians' and Teens' Perspective. *Malaysian Journal of Library & Information Science*, *18*(3), 75–90.

Abumandour, E.-S. T. (2020). Public Libraries' Role in Supporting E-learning and Spreading Lifelong Education: A Case Study. *Journal of Research in Innovative Teaching & Learning*.

Affelt, A. (2015). *The Accidental Data Scientist: Big Data Applications and Opportunities for Librarians and Information Professionals*. Medford: Information Today, Inc. https://doi.org/10.5596/c16-005

Ahmad, K., Zheng, J., & Muhammad, R. (2019). An Analysis of Academic Librarians Competencies and Skills for Implementation of Big Data Analytics in Libraries: A Correlational Study. *Data Technologies and Applications*.

Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of Big Data to Smart Cities. *Journal of Internet Services and Applications*, *6*(1), 25.

Anwyll, R., & Chawner, B. (2013). Social Media and Readers' Advisory A Win-Win Combination? *Reference & User Services Quarterly*, *53*(1), 18–22. https://doi.org/10.5860/rusq.53n1.18

Arnaboldi, M., Busco, C., & Cuganesan, S. (2017). Accounting, Accountability, Social Media and Big Data: Revolution or Hype? *Accounting, Auditing & Accountability Journal*, *30*(4), 762–776.

Azwar, M., & Sulthonah, S. (2018). The Utilization of Instagram as a Media Promotion: The Case Study of Library in Indonesia. *Insaniyat: Journal of Islam and Humanities*, *2*(2), 147–159.

Banica, L., & Hagiu, A. (2015). Big Data in Business Environment. *Scientific Bulletin-Economic Sciences*, *14*(1), 79–86.

Barriball, K. L., & While, A. (1994). Collecting Data Using a Semi-structured Interview: A Discussion Paper. *Journal of Advanced Nursing-Institutional Subscription*, *19*(2), 328–335.

Bellinger, G., Castro, D., & Mills, A. (2004). Data, Information, Knowledge,

and Wisdom. Retrieved from http://courseweb.ischool.illinois.edu/~katewill/spring2011-502/502 and other readings/bellinger on ackoff data info know wisdom.pdf

Bower, K. M. (2003). When to Use Fisher's Exact Test. In *American Society for Quality, Six Sigma Forum Magazine* (Vol. 2, pp. 35–37).

Bram, K., Bart-Jan, R., Scott, C., & Hans, de B. (2017). Big data in the Public Sector: Uncertainties and Readiness. *Information Systems Frontiers*, *19*(2), 267–283.

Brown, J. R. (1998). What is a Definition? *Foundations of Science*, *3*(1), 111–132.

Bryant, R., Katz, R. H., & Lazowska, E. D. (2008). Big-data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society. December.

Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big Data. *Business & Information Systems Engineering*, *5*(2), 65–69.

Bunskoek, K. (2014). 3 Key Hashtag Strategies: How to Market your Business and Content . Retrieved from https://blog.wishpond.com/post/62253333766/3-key-hashtag-strategies-how-to-market-your-business

Burgess, J., Galloway, A., & Sauter, T. (2015). Hashtag as Hybrid Forum: The Case of #agchatoz. In S. Jones (Ed.), *Hashtag publics. The Power and Politics of Discursive Networks.* (pp. 61–76). Pieterlen: Peter Lang.

Cahill, K. (2011). Going Social at Vancouver Public Library: What the Virtual Branch Did Next. *Program-Electronic Library and Information Systems*, *45*(3), 259–278. https://doi.org/10.1108/00330331111151584

Carlsson, H. (2012). Working with Facebook in Public Libraries: A Backstage Glimpse into the Library 2.0 Rhetoric. *Libri*, *62*(3), 199–210. https://doi.org/10.1515/libri-2012-0016

Cavanagh, M. F. (2015). Structuring an Action Net of Public Library Membership. *The Library Quarterly*, *85*(4), 406–426.

Cavanagh, M. F. (2016). Micro-blogging Practices in Canadian Public Libraries: A National Snapshot. *Journal of Librarianship and Information Science*, *48*(3), 247–259. https://doi.org/10.1177/0961000614566339

Chae, B. (2015). Insights from Hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for Supply Chain Practice and Research. *International Journal of Production Economics*, *165*, 247–259. https://doi.org/10.1016/j.ijpe.2014.12.037

Chang, C. C. (2018). Hakka Genealogical Migration Analysis Enhancement

Using Big Data on Library Services. *Library Hi Tech*, *36*(3), 426–442.

Chatten, Z., & Roughley, S. (2016). Developing Social Media to Engage and Connect at the University of Liverpool library. *New Review of Academic Librarianship*, *22*(2–3), 249–256.

Clement, J. (2019). Number of Monthly Active Instagram Users 2013-2018. Retrieved from https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/

Cooper, B. B. (2016). 10 Surprising New Twitter Stats to Help You Reach More Followers. Retrieved from https://blog.bufferapp.com/10-new-twitter-stats-twitter-statistics-to-help-you-reach-your-followers

Cox, M., & Ellsworth, D. (1997). Application-controlled Demand Paging for Out-of-core Visualization. In *Proceedings. Visualization'97 (Cat. No. 97CB36155)* (pp. 235–244). IEEE.

Crawford, S., & Syme, F. (2018). Enhancing Collection Development with Big Data Analytics. *Public Library Quarterly*, *37*(4), 387–393.

Cuong Nguyen, L., Partridge, H., & Edwards, S. L. (2012). Towards an Understanding of the Participatory Library. *Library Hi Tech*, *30*(2), 335–346.

Daer, A. R., Hoffman, R., & Goodman, S. (2014). Rhetorical Functions of Hashtag Forms across Social Media Applications. In *Proceedings of the 32nd ACM International Conference on The Design of Communication CD-ROM* (p. 16). ACM.

Dalla Valle, L., & Kenett, R. (2018). Social Media Big Data Integration: A New Approach Based on Calibration. *Expert Systems with Applications*, *111*, 76–90.

Dastrup, R. Adam. (2019). *Introduction to Human Geography*. Pressbooks. Retrieved from https://humangeography.pressbooks.com/

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is Big Data? A Consensual Definition and a Review of Key Research Topics. In *AIP Conference Proceedings* (Vol. 1644, pp. 97–104). AIP.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A Formal Definition of Big Data Based on its Essential Features. *Library Review*, *65*(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061 ER

de Zuniga, H. G., & Diehl, T. (2017). Citizenship, Social Media, and Big Data: Current and Future Research in the Social Sciences. *Social Science Computer Review*, *35*(1), 3–9.

Desouza, K. C., & Jacob, B. (2017). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration & Society*, *49*(7), 1043–1064.

Dewan, P. (2013). Reading Matters in the Academic Library. *Reference & User Services Quarterly*, *52*(4), 309–319. https://doi.org/10.5860/rusq.52n4.309

Dong, X. L., & Srivastava, D. (2013). Big Data Integration. In *2013 IEEE 29th Cnternational conference on Data Engineering (ICDE)* (pp. 1245–1248). IEEE.

Dranove, D. (2009). Empirical Methods in Strategy.

Dumbrell, D., & Steele, R. (2015). #worldhealthday 2014: The Anatomy of a Global Public Health Twitter Campaign. In T. X. Bui & R. H. Sprague (Eds.), *2015 48th Hawaii International Conference on System Sciences* (pp. 3094–3103). Los Alamitos: Ieee Computer Soc. https://doi.org/10.1109/hicss.2015.373

Dunn, K. (2000). Interviewing. Ch. 4, In, Hay, I. In I. Hay (Ed.), *Qualitative Research Methods in Human Geography*. Oxford University Press.

Durant, D. M., & Horava, T. (2015). The Future of Reading and Academic Libraries. *Portal: Libraries and the Academy*, *15*(1), 5–27.

Dwyer, N., & Marsh, S. (2014). What Can the Hashtag #trust Tell Us about How Users Conceptualise Trust? In *2014 Twelfth Annual International Conference on Privacy, Security and Trust (Pst)* (pp. 398–402).

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data Consumer Analytics and the Transformation of Marketing. *Journal of Business Research*, *69*(2), 897–904.

Evans, J. R., & Mathur, A. (2005). The Value of Online Surveys. *Internet Research*, *15*(2), 195–219.

Famoye, F., & Singh, K. P. (2006). Zero-inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. *Journal of Data Science*, *4*(1), 117–130.

Fan, W., & Gordon, M. D. (2014). Unveiling the Power of Social Media Analytics. *Communications of the ACM, in Press (June 2014)*, 26.

Fasola, O. S. (2015). Perceptions and Acceptance of Librarians towards Using Facebook and Twitter to Promote Library Services in Oyo State, Nigeria. *Electronic Library*, *33*(5), 870–882. https://doi.org/10.1108/el-04-2014-0066

Fernandez, J. (2009). A SWOT Analysis for Social Media in Libraries. *Online*, *33*(5), 35–37.

Fink, C., Schmidt, A., Barash, V., Cameron, C., & Macy, M. (2016). Complex Contagions and the Diffusion of Popular Twitter Hashtags in Nigeria. *Social Network Analysis and Mining*, *6*(1). https://doi.org/10.1007/s13278-015-0311-z

Finnish Ministry of Education and Culture. (2017). *Livrary Act*.

Fredriksson, C., Mubarak, F., Tuohimaa, M., & Zhan, M. (2017). Big Data in the Public Sector: A Systematic Literature Review. *Scandinavian Journal of Public Administration*, *21*(3), 39–62.

Fuller, M. (2015). Big Data: New Science, New Challenges, New Dialogical Opportunities. *Zygon®*, *50*(3), 569–582. https://doi.org/10.1111/zygo.12187

Gamage, P. (2016). New Development: Leveraging 'big data' Analytics in the Public Sector. *Public Money & Management*, *36*(5), 385–390.

Gan, C. M. (2016). A Survey of WeChat Application in Chinese Public Libraries. *Library Hi Tech*, *34*(4), 625–638. https://doi.org/10.1108/lht-06-2016-0068

Gandomi, A., & Haider, M. (2015). Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Ganster, L., & Schumacher, B. (2009). Expanding beyond Our Library Walls: Building an Active Online Community through Facebook. *Journal of Web Librarianship*, *3*(2), 111–128.

Garczynski, J. V. (2017). *Fundraising: How to Raise Money for Your Library Using Social Media*. Chandos Publishing.

Gartner. (2013). Big Data Defintion. *Gartner IT Glossary*. Retrieved from http://www.gartner.com/it-glossary/big-data/

Geurin-Eagleman, A. N., & Burch, L. M. (2016). Communicating via Photographs: A Gendered Analysis of Olympic Athletes' Visual Self-Presentation on Instagram. *Sport Management Review*, *19*(2), 133–145.

Giannoulakis, S., & Tsapatsoulis, N. (2015). Instagram Hashtags as Image Annotation Metadata. In R. Chbeir, Y. Manolopoulos, I. Maglogiannis, & R. Alhajj (Eds.), *Artificial Intelligence Applications and Innovations* (Vol. 458, pp. 206–220). Berlin: Springer-Verlag Berlin. https://doi.org/10.1007/978-3-319-23868-5_15

Gibbs, M., Meese, J., Arnold, M., Nansen, B., & Carter, M. (2015). #Funeral and Instagram: Death, Social Media, and Platform Vernacular. *Information Communication & Society*, *18*(3), 255–268. https://doi.org/10.1080/1369118x.2014.987152

Gilbert, C. J. H. E. (2014). Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Retrieved from http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

Goel, D., Palaniappan, S., & Arora, S. (2014). *Project TEXEMO*. Retrieved

from
https://projecttexemo.files.wordpress.com/2015/01/cmu_report.pdf

Gordon-Murnane, L. (2012). Big Data: A Big Opportunity for Librarians. *Online*, *36*(5), 30–34.

Gotti, F., Langlais, P., & Farzindar, A. (2014). Hashtag Occurrences, Layout and Translation: A Corpus-driven Analysis of Tweets Published by the Canadian Government. In *Lrec 2014 - Ninth International Conference on Language Resources and Evaluation* (pp. 2254–2261).

Guenduez, A. A., Mettler, T., & Schedler, K. (2020). Technological Frames in Public Administration: What Do Public Managers Think of Big Data? *Government Information Quarterly*, *37*(1), 101406.

Gul, R., & Ahsan, A. (2019). Big Data and Analytics: Case Study of Good Governance and Government Power. In *European Conference on Intangibles and Intellectual Capital* (pp. 128–XI). Academic Conferences International Limited.

Hall, H. (2011). Relationship and Role Transformations in Social Media Environments. *Electronic Library*, *29*(4), 421–428. https://doi.org/10.1108/02640471111156704

Hamed, A. A., & Wu, X. D. (2014). Does Social Media Big Data Make the World Smaller? An Exploratory Analysis of Keyword-Hashtag Networks. In C. Kesselman, P. Chen, & H. Jain (Eds.), *2014 Ieee International Congress on Big Data* (pp. 454–461). New York: Ieee. https://doi.org/10.1109/BigData.Congress.2014.72

Harkai, O. (2018). Instagram Marketing Practices 2018: Leveraging the Algorithm Change. Retrieved from https://blog.smarp.com/instagram-marketing-practices-in-2018-leveraging-the-algorithm-change

Harris, M. H. (1999). *History of Libraries of the Western World*. Scarecrow Press.

Harrison, A., Burress, R., Velasquez, S., & Schreiner, L. (2017). Social Media Use in Academic Libraries: a Phenomenological Study. *The Journal of Academic Librarianship*, *43*(3), 248–256.

Hicks, D., Cavanagh, M. F., & VanScoy, A. (2020). Social Network Analysis: A Methodological Approach for Understanding Public Libraries and Their Communities. *Library & Information Science Research*, *42*(3), 101029.

Hild, K. L. (2014). Outreach and Engagement through Instagram: Experiences with the Herman B Wells Library Account. *Indiana Libraries*, 30–32.

Holmberg, K., Huvila, I., Kronqvist-Berg, M., & Widén, G. (2009). What is

Library 2.0? *Journal of Documentation*, *65*(4), 668–681.

Holst, A. (2020). Information Created Globally 2010-2025.

Hopkins, P., Hare, J., Donaghey, J., & Abbott, W. (2015). Geo, Audio, Video, Photo: How Digital Convergence in Mobile Devices Facilitates Participatory Culture in Libraries. *The Australian Library Journal*, *64*(1), 11–22.

Hoy, M. B. (2014). Big Data: An Introduction for Librarians. *Medical Reference Services Quarterly*, *33*(3), 320–326. https://doi.org/10.1080/02763869.2014.925709

Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What We Instagram: A First Analysis of Instagram Photo Content and User Types. In *ICWSM*. University of Michigan.

Huang, H., Chu, S. K. W., & Chen, D. Y. T. (2015). Interactions Between English-Speaking and Chinese-Speaking Users and Librarians on Social Networking Sites. *Journal of the Association for Information Science and Technology*, *66*(6), 1150–1166. https://doi.org/10.1002/asi.23251

Hussain, A. (2015). Adoption of Web 2.0 in Library Associations in the Presence of Social Media. *Program-Electronic Library and Information Systems*, *49*(2), 151–169. https://doi.org/10.1108/prog-02-2013-0007

Huvila, I., Holmberg, K., Kronqvist-Berg, M., Nivakoski, O., & Widén, G. (2013). What Is Librarian 2.0–New Competencies or Interactive Relations? A Library Professional Viewpoint. *Journal of Librarianship and Information Science*, 0961000613477122.

Huwe, T. K. (2014). Big Data and the Library: A Natural Fit. *Computers in Libraries*, *34*(2), 17–18.

Ibrahim, Y. (2015). Instagramming Life: Banal Imaging and the Poetics of the Everyday. *Journal of Media Practice*, *16*(1), 42–54.

Islam, M. M., & Habiba, U. (2015). Use of Social Media in Marketing of Library and Information Services in Bangladesh. *Desidoc Journal of Library & Information Technology*, *35*(4), 299–303.

Jaeger, P. T., Greene, N. N., Bertot, J. C., Perkins, N., & Wahl, E. E. (2012). The Co-evolution of E-government and Public Libraries: Technologies, Access, Education, and Partnerships. *Library & Information Science Research*, *34*(4), 271–281.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European conference on machine learning* (pp. 137–142). Springer.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, Challenges,

Tools and Good Practices. In *2013 Sixth international conference on contemporary computing (IC3)* (pp. 404–409). IEEE.

Kaul, H. K. (2016). Libraries and the Social Media Networks. *Desidoc Journal of Library & Information Technology*, *36*(5), 257–260.

Kaushik, A. (2016). Use of Social Networking Sites Tools and Services by LIS Professionals for Libraries: A Survey. *Desidoc Journal of Library & Information Technology*, *36*(5), 284–290.

Kho, N. D. (2018). The State of Big Data. *Econtent*, *41*(1), 11–12.

Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data Applications in the Government Sector. *Communications of the ACM*, *57*(3), 78–85.

Kim, T.-Y., Gang, J.-Y., & Oh, H.-J. (2019). Spatial Usage Analysis Based on User Activity Big Data Logs in Library. *Library Hi Tech*.

Kim, Y.-S., & Cooke, L. (2017). Big Data Analysis of Public Library Operations and Services by Using the Chernoff Face Method. *Journal of Documentation*, *73*(3), 466–480. https://doi.org/10.1108/JD-08-2016-0098 ER

Kirjastot.fi. (n.d.). Finnish Public Libraries Statistics.

Kitchin, R., & McArdle, G. (2016). What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society*, *3*(1), 2053951716631130.

Koltay, T. (2016). Library and Information Science and the Digital Humanities: Perceived and Real Strengths and Weaknesses. *Journal of Documentation*, *72*(4), 781–792.

Krawczyk, D. C. (2018). Chapter 9 - Deduction and Induction. In D. C. B. T.-R. Krawczyk (Ed.), *Reasoning: The Neuroscience of How We Think* (pp. 199–225). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-809285-9.00009-0

Kronqvist-Berg, M. (2014). *Social Media and Public Library: Exploring Information Activities of Library Professionals and Users*. *Information Studies*. Åbo Akademi, Turku, Finland.

Kumar, V., & Singh, A. P. (2015). Impact of Information Explosion on Library Professionals in Digital Technology Scenario. *Transforming Dimension of IPR: Challenges for New Age Libraries*, 530.

Kuoppakangas, P., Kinder, T., Stenvall, J., Laitinen, I., Ruuskanen, O.-P., & Rannisto, P.-H. (2019). Examining the Core Dilemmas Hindering Big Data-related Transformations in Public-Sector Organisations. *NISPAcee Journal of Public Administration and Policy*, *12*(2), 131–156.

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine Learning with Big Data: Challenges and Approaches. *IEEE*

*Access*, *5*, 7776–7797.

La Sorte, F. A., Lepczyk, C. A., Burnett, J. L., Hurlbert, A. H., Tingley, M. W., & Zuckerberg, B. (2018). Opportunities and Challenges for Big Data Ornithology. *The Condor: Ornithological Applications*, *120*(2), 414–426.

Lamont, L., & Nielsen, J. (2015). Calculating Value: A Digital Library's Social Media Campaign. *Bottom Line*, *28*(4), 106–111. https://doi.org/10.1108/bl-07-2015-0010

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note*, *6*(February 2001), 1–4.

Lee, I., & Lee, K. (2015). The Internet of Things (IoT): Applications, Investments, and Challenges for Enterprises. *Business Horizons*, *58*(4), 431–440.

Lee, K. (2016). How to Use Hashtags: How Many, Best Ones, and Where to Use Them. Retrieved from https://blog.bufferapp.com/a-scientific-guide-to-hashtags-which-ones-work-when-and-how-many

LGMA. (2012). *Public Library Authority Statistics Actuals 2011*. Retrieved from http://www.askaboutireland.ie/aai-files/assets/libraries/an-chomhairle-leabharlanna/libraries/public-libraries/publications/2011-library-statistics-actuals.pdf

Li, J., Tao, F., Cheng, Y., & Zhao, L. (2015). Big Data in Product Lifecycle Management. *The International Journal of Advanced Manufacturing Technology*, *81*(1–4), 667–684.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Louis, M. R. (1980). Surprise and Sense Making: What Newcomers Experience in Entering Unfamiliar Organizational Settings. *Administrative Science Quarterly*, 226–251.

Lyman, P., & Varian, H. R. (2000). Reprint: How Much Information? *Journal of Electronic Publishing*, *6*(2).

Ma, Z., Sun, A., & Cong, G. (2013). On Predicting The Popularity Of Newly Emerging Hashtags in Twitter. *Journal of the American Society for Information Science and Technology*, *64*(7), 1399–1410.

Maciejewski, M. (2017). To Do More, Better, Faster and More Cheaply: Using Big Data in Public Administration. *International Review of Administrative Sciences*, *83*(1_suppl), 120–135.

Mahraj, K. (2012). Reference Services Review: Content Analysis, 2006-2011. *Reference Services Review*, *40*(2), 182–198.

Malomo, F., & Sena, V. (2017). Data Intelligence for Local Government? Assessing the Benefits and Barriers to Use of Big Data in the Public

Sector. *Policy & Internet*, *9*(1), 7–27.

Mandel, L. H. (2013). Finding Their Way: How Public Library Users Wayfind. *Library & Information Science Research*, *35*(4), 264–271.

Manovich, L. (2016). *Instagram and Contemporary Image*. *Manovich. net, New York*.

Mapulanga, P. (2013). Digitising Library Resources and Building Digital Repositories in the University of Malawi Libraries. *Electronic Library*, *31*(5), 635–647.

Marsland, S. (2014). *Machine Learning: An Algorithmic Perspective*. Chapman and Hall/CRC.

Martynas Mažvydas National Library of Lithuania. (2015). *General 2015 Library Report*. Retrieved from https://www2.lnb.lt/media/public/english/pdf/EN_2015_Library_Statistics.pdf

McNely, B. J. (2012). Shaping Organizational Image-power through Images: Case Histories of Instagram. In *2012 IEEE International Professional Communication Conference* (pp. 1–8). IEEE.

Merhi, M. I., & Bregu, K. (2020). Effective and Efficient Usage of Big Data Analytics in Public Sector. *Transforming Government: People, Process and Policy*.

Moise, I. (2016). The Technical Hashtag in Twitter Data: a Hadoop Experience. In *2016 Ieee International Conference on Big Data (Big Data)* (pp. 3519–3528).

Mon, L., & Lee, J. (2015). Influence, Reciprocity, Participation, and Visibility: Assessing the Social Library on Twitter. *Canadian Journal of Information and Library Science-Revue Canadienne Des Sciences De L Information Et De Bibliotheconomie*, *39*(3–4), 279–294.

Natioanl Statistics Office. (n.d.). Statistics on Libraries.

Noh, Y. (2015). Imagining Library 4.0: Creating a Model for Future Libraries. *Journal of Academic Librarianship*, *41*(6), 786–797. https://doi.org/10.1016/j.acalib.2015.08.020

Oh, C., Lee, T., Kim, Y., Park, S., & Suh, B. (2016). Understanding Participatory Hashtag Practices on Instagram: A Case Study of Weekend Hashtag Project. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1280–1287). ACM.

Okwechime, E., Duncan, P., & Edgar, D. (2018). Big Data and Smart Cities: A Public Sector Organizational Learning Perspective. *Information Systems and E-Business Management*, *16*(3), 601–625.

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic

researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology*, *8*(5), 375–387.

Ortiz-Ospina, E. (2019). The Rise of Social Media. Retrieved August 15, 2020, from https://ourworldindata.org/rise-of-social-media

Owais, S. S., & Hussein, N. S. (2016). Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. *International Journal of Advanced Computer Science and Applications*, *7*(3), 254–258.

Patel, B., Roy, S., Bhattacharyya, D., & Kim, T.-H. (2017). Necessity of Big Data and Analytics for Good e-governance. *International Journal of Grid and Distributed Computing*, *10*(8), 11–20.

Peña-López, I. (2005). ITU Internet Report 2005: The Internet of Things.

Phillips, N. K. (2011). Academic Library Use of Facebook: Building Relationships with Students. *The Journal of Academic Librarianship*, *37*(6), 512–522.

Pinho-Costa, L., Yakubu, K., Hoedebecke, K., Laranjo, L., Reichel, C. P., Colon-Gonzalez, M. D., … Errami, H. (2016). Healthcare Hashtag Index Development: Identifying Global Impact in Social Media. *Journal of Biomedical Informatics*, *63*, 390–399. https://doi.org/10.1016/j.jbi.2016.09.010

Pittman, M., & Reich, B. (2016). Social Media and Loneliness: Why an Instagram Picture May Be Worth More Than a Thousand Twitter Words. *Computers in Human Behavior*, *62*, 155–167.

Prior, G., Toombs, B., Taylor, L., & Currenti, R. (2013). *Cross-European Survey to Measure Users' Perceptions of the Benefits of ICT in Public Libraries*. Bill & Melinda Gates Foundation.

Python. (n.d.). Python Language Reference. Retrieved from http://www.python.org

Queirós, A., Faria, D., & Almeida, F. (2017). Strengths and limitations of qualitative and quantitative research methods. *European Journal of Education Studies*.

Rajaraman, V. (2016). Big Data Analytics. *Resonance*, *21*(8), 695–716.

Rasmussen, C. H. (2016). The Participatory Public Library: The Nordic Experience. *New Library World*.

Reinhalter, L., & Wittmann, R. J. (2014). The Library: Big Data's Boomtown. *Serials Librarian*, *67*(4), 363–372. https://doi.org/10.1080/0361526X.2014.915605

Republic of Bulgaria National Statistical Institute. (n.d.). Euro-SDMX Metadata Structure (ESMS).

Rich, J., Haddadi, H., Hospedales, T. M., & Acm. (2016). Towards Bottom-Up Analysis of Social Food. *Dh'16: Proceedings of the 2016 Digital Health Conference*, 111–120. https://doi.org/10.1145/2896338.2897734

Robinson, L. (2009). Information Science: Communication Chain and Domain Analysis. *Journal of Documentation*, *65*(4), 578–591.

Rowley, J. (2007). The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science*, *33*(2), 163–180.

Saggi, M. K., & Jain, S. (2018). A Survey towards an Integration of Big Data Analytics to Big Insights for Value-creation. *Information Processing & Management*, *54*(5), 758–790.

Salomon, D. (2013). Moving on from Facebook: Using Instagram to Connect with Undergraduates and Engage in Teaching and Learning. *College & Research Libraries News*, *74*(8), 408–412.

Sandelius, N. (2012). Imanging and Explore-Finnish Libraries Now!

Sarker, M. N. I., Wu, M., & Hossin, M. A. (2018). Smart Governance through Big Data: Digital Transformation of Public Agencies. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 62–70). IEEE.

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The Measurement of Work Engagement with a Short Questionnaire: A Cross-national Study. *Educational and Psychological Measurement*, *66*(4), 701–716.

Scott, A. J., & Holt, D. (1982). The Effect of Two-stage Sampling on Ordinary Least Squares Methods. *Journal of the American Statistical Association*, *77*(380), 848–854.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion Knowledge: Further Exploration of a Prototype Approach. *Journal of Personality and Social Psychology*, *52*(6), 1061.

Shcherbina, I. (2017). Reading in the Age of Web 2.0. *Russian Social Science Review*, *58*(1), 86–108.

Sheldon, P., & Bryant, K. (2016). Instagram: Motives for Its Use and Relationship to Narcissism and Contextual Age. *Computers in Human Behavior*, *58*, 89–97.

Siapera, E. (2014). Tweeting #Palestine: Twitter and the Mediation of Palestine. *International Journal of Cultural Studies*, *17*(6), 539–555. https://doi.org/10.1177/1367877913503865

Sison, R., & Shimura, M. (1998). Student Modeling and Machine Learning. *International Journal of Artificial Intelligence in Education (IJAIED)*, *9*, 128–158.

Small, T. A. (2011). What the Hasgtag? A Content Analysis of Canadian Politics on Twitter. *Information Communication & Society*, *14*(6), 872–

895. https://doi.org/10.1080/1369118x.2011.554572

Stathopoulou, A., Borel, L., Christodoulides, G., & West, D. (2017). Consumer Branded #Hashtag Engagement: Can Creativity in TV Advertising Influence Hashtag Engagement? *Psychology & Marketing*, *34*(4), 448–462. https://doi.org/10.1002/mar.20999

Statistics Denmark. (n.d.). BIB1: Public Libraries Key Figures by Region and Key Figures.

Stejskal, J., & Hajek, P. (2015). Effectiveness of Digital Library Services as a Basis for Decision-making in Public Organizations. *Library & Information Science Research*, *37*(4), 346–352.

Storey, V. C., & Song, I.-Y. (2017). Big Data Technologies and Management: What Conceptual Modeling Can Do. *Data & Knowledge Engineering*, *108*, 50–67.

Stuart, E., Stuart, D., & Thelwall, M. (2017). An Investigation of the Online Presence of UK Universities on Instagram. *Online Information Review*, *41*(5), 582–597.

Surbakti, F. P. S., Wang, W., Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. *Information & Management*, *57*(1), 103146.

Tekulve, N., & Kelly, K. (2013). Worth 1,000 words: Using Instagram to engage library users.

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *The 8th International AAAI Conference on Weblogs and Social Media*. Oxford.

Vakkari, P., Aabø, S., Audunson, R., Huysmans, F., & Oomes, M. (2014). Perceived Outcomes of Public Libraries in Finland, Norway and the Netherlands. *Journal of Documentation*, *70*(5), 927–944.

Vakkari, P., & Serola, S. (2012). Perceived Outcomes of Public Libraries. *Library & Information Science Research*, *34*(1), 37–44.

Vanwynsberghe, H., Boudry, E., Vanderlinde, R., & Verdegem, P. (2014). Experts as Facilitators for the Implementation of Social Media in the Library? A Social Network Approach. *Library Hi Tech*, *32*(3), 529–545. https://doi.org/10.1108/lht-02-2014-0015

Vanwynsberghe, H., Vanderlinde, R., Georges, A., & Verdegem, P. (2015). The Librarian 2.0: Identifying a Typology of Librarians' Social Media Literacy. *Journal of Librarianship and Information Science*, *47*(4), 283–293. https://doi.org/10.1177/0961000613520027

Vassilakaki, E., & Garoufallou, E. (2014). The Impact of Facebook on Libraries and Librarians: A Review of the Literature. *Program-*

*Electronic Library and Information Systems*, *48*(3), 226–245. https://doi.org/10.1108/prog-03-2013-0011

Vassilakaki, E., & Garoufallou, E. (2015). The Impact of Twitter on Libraries: A Critical Review of the Literature. *Electronic Library*, *33*(4), 795–809. https://doi.org/10.1108/el-03-2014-0051

Vydra, S., & Klievink, B. (2019). Techno-optimism and Policy-pessimism in the Public Sector Big Data Debate. *Government Information Quarterly*, *36*(4), 101383.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'Big Data' Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study. *International Journal of Production Economics*, *165*, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

Wang, C., Wang, Q., Ren, K., & Lou, W. (2010). Privacy-preserving Public Auditing for Data Storage Security in Cloud Computing. In *INFOCOM, 2010 Proceedings IEEE* (pp. 1–9). Ieee.

Wang, R., Liu, W. L., & Gao, S. Y. (2016). Hashtags and Information Virality in Networked Social Movement Examining Hashtag Co-Occurrence Patterns. *Online Information Review*, *40*(7), 850–866. https://doi.org/10.1108/oir-12-2015-0378

Waxman, L., Clemons, S., Banning, J., & McKelfresh, D. (2007). The library as Place: Providing Students with Opportunities for Socialization, Relaxation, and Restoration. *New Library World*, *108*(9/10), 424–434.

West, C. (2019). 17 Instagram Stats Marketers Need to Know for 2019. Retrieved November 4, 2019, from https://sproutsocial.com/insights/instagram-stats/

Widén, G., Soumi, R., Joas, M., Zhan, M., Fredriksson, C., Tuohimaa, M., & Mubarak, F. (2016). Big Cities Meet Big Data. Turku, Finland. Retrieved from http://blogs2.abo.fi/bcbd/

Wilkinson, J. (2018). Accessible, Dynamic Web Content Using Instagram. *Information Technology and Libraries*, *37*(1), 19–26.

Wojcik, M. (2015). The Use of Web 2.0 Services by Urban Public Libraries in Poland: Changes over the Years 2011-2013. *Libri*, *65*(2), 91–103. https://doi.org/10.1515/libri-2015-0017

Xie, I., & Stevenson, J. (2014). Social Media Application in Digital Libraries. *Online Information Review*, *38*(4), 502–523. https://doi.org/10.1108/oir-11-2013-0261

Xu, M., Cai, H., & Liang, S. (2015). Big data and Industrial Ecology. *Journal of Industrial Ecology*, *19*(2), 205–210.

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing*

*Surveys (CSUR)*, *50*(2), 25.

Yigitcanlar, T., Desouza, K. C., Butler, L., & Roozkhosh, F. (2020). Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature. *Energies*, *13*(6), 1473.

Yin, S., & Kaynak, O. (2015). Big data for Modern Industry: Challenges and Trends. *Proceedings of the IEEE*, *103*(2), 143–146.

Ylipulli, J., & Luusua, A. (2019). Without Libraries What Have We? Public Libraries As Nodes For Technological Empowerment In The Era Of Smart Cities, AI And Big Dataa. In *Proceedings of the 9th International Conference on Communities & Technologies-Transforming Communities* (pp. 92–101).

Young, Scott W H. (2016). Introduction to Social Media Optimization: Setting the Foundation for Building Community. *Library Technology Reports*, *52*(8), 5–8.

Young, Scott Woodward Hazard, & Rossmann, D. (2015). Building Library Community through Social Media. *Information Technology and Libraries*, *34*(1), 20–37.

Zappavigna, M. (2015). Searchable Talk: The Linguistic Functions of Hashtags. *Social Semiotics*, *25*(3), 274–291. https://doi.org/10.1080/10350330.2014.996948

Zappavigna, Michele. (2016). Social Media Photography: Construing Subjectivity in Instagram Images. *Visual Communication*, *15*(3), 271–292.

Zhan, M., & Widén, G. (2018). Public Libraries: Roles in Big Data. *The Electronic Library*, *36*(1), 133–145.

Zhong, Y., Zhang, H., & Jain, A. K. (2000). Automatic Caption Localization in Compressed Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(4), 385–392.

Zou, H., Chen, H. M., & Dey, S. (2015). Exploring User Engagement Strategies and Their Impacts with Social Media Mining: The Case of Public Libraries. *Journal of Management Analytics*, *2*(4), 295–313.

Zyl, A. S. Van. (2009). The Impact of Social Networking 2.0 on Organisations. *Electronic Library*, *27*(6), 906–918. https://doi.org/10.1108/02640470911004020

ISBN 978-952-12-4070-6