School of Media, Creative Arts and Social Inquiry

Analysing the Usage Data of Open Access Scholarly Books: What Can Data Tell Us?

Alkim Ozaygen

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

July 2019

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material, which has been accepted for the award of any other degree or diploma in any university.

Signature:

Date:

Abstract

Despite the vast amount of research available on open access (OA) journal articles, little is known about OA monographs—scholarly books written on a single subject—which constitute the major means of communicating research and results in the humanities and social sciences (HSS).

Unlike that for journal articles, the OA monograph market is diversified and usually dominated by university-based collaborations. Different publishers are experimenting with different business models (Adema, 2010). Moreover, in addition to length, OA monographs differ from OA journal articles in many respects: they can reside in multiple repositories, be in a multitude of file formats, and have different types and numbers of identifiers. Thus, patterns of user interactions, including usage, access and mentions, may differ across OA journal articles and OA monographs. In addition, these differences make the data relating to OA monographs much more difficult to capture.

A study capturing, analysing and interpreting data related to the usage of monographs is not only needed to fill this gap in scholarly knowledge, but also because major gaps exist in knowledge about OA books held by various parties in the monograph industry, including publishers, repositories, libraries and research funders. The findings of such a study would help these parties understand how their titles are used and disseminated across the Internet from the point that they are made OA.

This study aims to fill this gap by exploring data related to the 28 titles made OA during the first pilot phase of Knowledge Unlatched (KU). These 28 titles were part of the proof-of-concept for KU's new approach to creating a sustainable route to OA for scholarly books. The data discussed in this thesis reflect different types of digital acts, for example, downloading, sharing, mentioning and citing, related to the 28 KU pilot collection books. The study captures and examines these digital acts on OA books using various approaches, including webometrics and altmetrics methods. Subsequently, it explores the relations between the acts that occur on social media and interprets these findings using citations, personal behaviour and social theories.

To uncover the factors that impact on the digital uses of OA books, it explores relationships between book characteristics and the characteristics and motivations of the groups using and sharing books in digital landscapes. Finally, this data is used to propose a causal chain model of user acts and motivations, which has the potential to be applied more widely and to other types of research output.

Dedication

This dissertation is dedicated to Nesrin Güner, who supported me during my hard times. Thank you for everything and I miss you!

Acknowledgements

First of all, I would like to thank John Hartley who made my study at Curtin University possible, and in so doing allowed me to become part of the CCAT family. I would like to express my deepest gratitude to Lucy Montgomery for her support, guidance and patience throughout my research. She introduced me to the world of OA monographs and the many projects related to books taking place within this global community. I am deeply indebted to Tama Leaver for his guidance and trust. The opportunity to work on the Tracking Infrastructure for Social Media Analysis (TrISMA) project while undertaking my PhD proved to be invaluable. I would also like to extend my deepest gratitude to Cameron Neylon who provided many important lessons on how to approach metrics analysis, and on how to structure a thesis. Cameron's willingness to come in to the university to discuss my thesis with me during his Christmas break meant a great deal, and I will be forever grateful.

I would like to thank Frances Pinter for her support, feedback and encouragement during my PhD. I had the privilege of working with Frances on a number of projects relating to OA books during the course of my study – and Frances's generosity as a colleague and mentor cannot be overstated. Climbing the Acropolis with Frances in Athens was a particular highlight!

I would like to thank Eelco Ferwerda and Ronald Snijder from OAPEN for their willingness to share data relevant to my research, and their genuine passion for OA books and the communities that read, write and use them. Similarly, the HathiTrust Foundation's willingness to make data available for research purposes added enormously to my capacity to carry out this project.

Sonia Dickinson from Curtin University School of Marketing provided me with access to use the Radian6 software. In so doing, she made it possible for me to gather the social network data that much of this thesis builds on.

My thanks also go to Pierre Mounier from OpenEdition; Charles Watkinson and Rebecca Walzenbach from Michigan University Press; Lara Speicher and Alison

Fox from UCL Press, and Rupert Gatti from Open Book Publishers for their willingness to help me to identify data and perspectives relevant to this project.

Tableau's Academic Program provided me with the free license that allowed me to develop the first prototype dashboards for the KU pilot collections, and to gather feedback from publishers about the dashboards and their usefulness.

Thanks also go to Shanshan Liu for being my housemate and friend. Together, we shared our great PhD journey. Margaret Colyer shared her home – and made me feel at home - for my first two years in Australia. Thanks go to Ben, Eddie for helping me and making me feel at home. I will always remember Romit Dasgupta for being one of my first friends in Perth and introducing me to the Turkish community here.

A big thanks to Burcu Şimşek for her motivation and support, and also for introducing me to John Hartley and opening the door to study here in Australia.

I also want to thank Fazıl Gökgöz for supporting me during my PhD journey. Henry Silling Li and Ali Mozaffari for providing me my first orientation to school when I first came to Curtin University back in 2014 to do my research.

My friends Haldun, Kerem, Ayşegül, Nar, Fatih, Uluç and Hakan for being with me during this journey.

Altay from whom I first heard R programming language. To my parents Meral and Tuna for always supporting me in any decision I took all my life.

Lastly, the most important thanks go to my wife Gül and step daughter Ekin for being with me in Perth, and for showing their support and love. Having Gül and Ekin with my in Perth during the final two years of PhD study made all the difference. Thank you for coming with me on this journey.

Table of contents

List of Figures

List of Tables

List of Abbreviations

ACLS HEB	American Council of Learned Societies Humanities E-Book Project
ALM	Article Level Metrics
API	Application Programming Interface
ARWU	Academic Ranking of World Universities
BASE	Bielefeld Academic Search Engine
BKCI	Thomson Reuters Book Citation Index
CAUL	The Council of Australian University Librarians
CNRS	Centre national de la recherche scientifique. The French National Centre for Scientific Research
DMI	Digital Methods Initiative
DOAB	The Directory of Open Access Books
DOI	Digital Object Identifier
EC	European Commission
EDS	EBSCO Discovery Service
ERC	European Research Council
F1000	Faculty of 1000
HAL	Hyper Articles en Ligne.
HEFCE	The Higher Education Funding Council for England
HIRMEOS	High Integration of Research Monographs in the European Open Science
HSS	Humanities and social sciences
HTML	Hypertext Markup Language

HTTP	Hypertext Transfer Protocol
ICT	Information and Communication Technologies
IQR	The interquartile range
IRUS-UK	Institutional Repository Usage Statistics UK
ISBN	International Standard Book Number
ISI	Institute for Scientific Information
JMIR	The Journal of Medical Internet Research
JSTOR	Journal Storage
KU	Knowledge Unlatched
NCBI	National Centre for Biotechnology Information
NIH	National Institute of Health
NISO	The National Information Standards Organization
NLM	United States National Library of Medicine
NWO	Nederlandse Organisatie voor Wetenschappelijk Onderzoek. The Netherlands Organisation for Scientific Research
OA	Open Access
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAPEN	Open Access Publishing in European Networks
OpenAIRE	Open access infrastructure for research in Europe
OpenDOAR	Directory of Open Access Repositories
OPERAS	Open Access in the European Research Area Through Scholarly Communication
PDF	Portable Document Format
PLOS	The Public Library of Science

ROAR	Registry of Open Access Repositories
SLD	Second-Level Domain
TCAT	Twitter Capture Tool
TLD	Top-Level Domain
UCL	University College London
UNESCO	The United Nations Educational, Scientific and Cultural Organization
URL	Uniform Resource Locator
WoS	Web of Science
XML	eXtensible Markup Language

1 Introduction

With the digitisation of scholarly communication and the rise of social media, online communication has come to play an ever more important role in research and knowledge sharing. Journal articles and monographs that were previously only printed are now published digitally as well. Many of these scholarly publications are also open access (OA): available at no cost to everyone who has access to the Internet. Scholars now discuss research on social networks, including Twitter and Facebook. Many use online platforms and software, which provide them with tools to download, store, annotate, discuss, and share research outputs, and even write their general articles on their blogs (Mas-Bleda, et al., 2014). The vast number of interactions of users with digital platforms has created an environment rich with data. It is now possible to interrogate data about how scholars communicate and to explore questions that could not have been answered in an analogue world.

In this rich data landscape, it is now possible to extract new insights on different aspects of interactions. Among the conventional scientific research outputs, monographs have not been the focus of studies interested in capturing and examining interactions with and around publications. A monograph is defined as a specialist scholarly book, usually written by a single author on a single subject. Unlike a textbook, which surveys the state of knowledge in a field, the main purpose of a monograph is to present primary research and original scholarship. Monographs are most commonly produced within the humanities and social sciences (HSS), although scholars within the hard sciences also publish and use them. Due to a number of factors, the monograph market accepted the transition to OA later than academic journals. The monograph market has also evolved differently, leading it to be positioned differently in the academic publishing industry. The goal of this study is to fill the gap in our knowledge relating to OA monographs in scholarly communication. Using an extensive set of data, it identifies and exposes the dynamics of these interactions in order to shed light on how and why OA monographs are discovered, accessed, used, mentioned, cited and receive ratings

from book platforms. These dynamics would help us to understand how these titles are disseminated in the digital landscape.

1.1 The Monograph Market

After the 1980s, global monograph sales experienced a period of dramatic decline (Ryan et al., 2002; Steele, 2008; Willinsky, 2009). According to figures provided by the co-directors of the ACLS Humanities E-Book (HEB) Project[1], average per-title sales in the monograph market have plummeted, from about 2,000 copies in 1980, to 1,000 in the late 1980s, to 500 by 1990 (Gardiner & Musto, 2005). This is mainly due to monographs' high cost and the decreasing proportion of library budgets available to support monograph purchases. Today, monograph print runs often comprise just 200–300 copies in hardback, with cover prices ranging from US$50 to US$250 per copy to cover publishers' costs. Rather than being sold to individual readers, the main markets for monographs tend to be well-funded university libraries in Western Europe and the United States (Montgomery, 2014). As journal prices have continued to increase, library spending on monographs has come under pressure. In a survey of 109 librarians in UK universities conducted by OAPEN-UK in 2014 (OAPEN-UK, 2014), 42% said that they usually do not have enough money to buy all of the monographs they need each year (Collins & Milloy, 2016).

Digitization and Internet technologies have changed the production, distribution and consumption patterns associated with academic publishing. These technologies enable more open, transparent, and diverse approaches to scholarly communication. There is a growing trend towards OA in both journals and monographs. Research materials and original datasets are increasingly being shared online in both open and closed formats, and research collaborations now often take place over digital networks.

[1] ACLS Humanities E-Book (HEB) project is an online, fully searchable collection of nearly 3,000 books of major importance in the humanities. Titles are chosen in collaboration with 20 ACLS member societies and over 100 publishers. Originally funded as the ACLS History E-Book Project in 1999 by a $3-million, five-year grant from The Andrew W. Mellon Foundation, with additional funding from the Gladys Krieble Delmas Foundation, HEB achieved self-sustainability in 2005 (https://www.humanitiesebook.org/).

The adoption of OA mandates by research funders has been associated with the emergence of a number of initiatives focused on the use of new technologies and OA to improve the dissemination and expand the readership of scholarly monographs. A key European initiative in this space is Open Access Publishing in European Networks (OAPEN), a project initially funded by the European Commission, whose objective is to promote OA book publishing and increase the visibility of and access to high-quality book-length scholarly publications. In keeping with its vision of both national-level experimentation and testing and international collaboration among OA monograph projects and stakeholders, OAPEN has developed two national satellite projects: OAPEN-UK and OAPEN-NL (Eve, 2014). These national-level projects are focused on gathering data and carrying out research on OA monographs ("Background to OAPEN-UK," n.d.). The OAPEN repository (https://oapen.org/)provides hosting, preservation and discovery services for more than 2,000 OA academic books. These include a collection of books made OA by the not-for-profit initiative Knowledge Unlatched (KU), which this research focuses on.

Knowledge Unlatched was established in 2012 to coordinate library support and funding for OA scholarly books. Knowledge Unlatched is not a publisher. Rather, its model is based on libraries around the world sharing a fixed 'title fee', which is paid to publishers in return for making books available with an OA license. In October 2013, KU launched its pilot collection, consisting of 28 new titles in the fields of anthropology, history, literature, media and communications, and politics provided by 13 recognised scholarly publishers. Almost 300 libraries from 24 countries agreed to contribute to the costs of making the pilot collection OA. The KU pilot collection titles were originally hosted on three platforms (OAPEN, HathiTrust, and Internet Archive). Between March 2014 and June 2015, nearly 36,000 downloads from 167 countries were reported for the 28 books in the pilot collection (Emery, 2015). In March 2017, KU announced that their second collection, consisting of an additional 78 new books from 26 publishers, had been successfully made open access. Knowledge Unlatched continues to make additional titles OA as the model continues to scale. This study is based on data relating to KU pilot collection titles published in 2013 and 2014 for the period between March 2014 and July 2017.

1.2 Capturing digital traces of scholarly outputs

Digitization and Internet technologies are not only resulting in new publishing and dissemination models. They are also creating new possibilities for capturing data relating to the visibility and use of research outputs, as well as new challenges for scholarly communication stakeholders.

Several approaches are used to capture data on the visibility and usage of research outputs. In an effort to draw attention to the need for filters capable of encompassing online activities, Jason Priem coined the term 'altmetrics', short for 'alternative metrics', in 2010 (Priem, 2010; Priem, Taraborelli, Groth, & Neylon, 2010). Altmetrics has become a popular approach to capturing data on how research outputs are being shared, used and discussed on social media and publishers' sites. Although altmetrics methodologies have been applied to collect data on journal articles, very little research has applied these techniques to specialist scholarly books.

Long before altmetrics, in 1997, the concept of webometrics was introduced (Almind & Ingwersen, 1997; Bornmann, 2014a). Almind and Inwersen applied the idea of conducting statistical analyses of scientific journal citation patterns to web-based content. This approach mainly uses quantitative analysis to identify the number and types of hyperlinks. One of the uses of webometrics is to assess the impact of ideas or documents on the web by analyzing their online presence (Thelwall, 2009).

1.3 Issues related to monographs

Monographs have not been a focus in the development of altmetrics for a number of reasons. Primary amongst these is the fact that books have been much slower to shift to digital and online formats than journals, limiting the scope for tracking monographs across digital landscapes. The challenges associated with usage data for monographs differ in important ways from those associated with journal articles. Firstly, there is more diversity and scarcity in the monograph publishing market. Unlike the scientific journal market, which is dominated by a few big publishers, many publishers of different sizes operate on the OA monograph market. Small publishers of OA monographs find it almost impossible to analyze altmetric data

associated with their publications, as they generally lack the resources to host in-house teams with the capacity to handle complex data from multiple sources. Secondly, OA monographs are often hosted in more than one repository, which makes it more difficult to collect comprehensive data about their usage. Another problem are the diverse formats in which monographs are available, which include PDF, ePUB, HTML and Kindle MOBI. In addition to these formats, monographs can be divided into chapters, as is the case with the digital library JSTOR, which hosts OA books in the form of chapter-level downloads. These issues pose challenges in tracking monograph usage. For example, unlike journal articles, which have a single Digital Object Identifier (DOI), an OA monograph can have a number of different identifiers. First, it can have commercial book identifiers, the International Standard Book Number (ISBN) for each format in which it is available, including hardcover, softcover, PDF, ePUB and HTML. In addition to ISBN identifiers, the monograph may also have a separate DOI identifier for each of these formats assigned by each repository. Thus, to investigate a monograph's mentions on social media, it is not possible to query a single DOI, as is the case with published journal articles, with the exception of their preprints, which can be hosted in a different repository and are assigned a separate DOI.

1.4 Research Questions and Significance

The main research question of this study is What can rich data reveal about the use of and interactions related to Open Access monographs?

The project draws on data relating to 28 OA monographs made available by KU for the period between March 2014 and July 2017. An additional title from UCL Press is used for the network analysis on Twitter.

This study includes three distinct research phases:

1. Exploring the extent to which the use of and interactions relating to KU OA books can be detected across global digital landscapes;

2. Identifying and investigating the relationship between these interactions; and, finally:

3. Interpreting the detected interactions using social theories and citation theories and identifying the factors that affect them.

This study is the first to focus on capturing data related to the discovery, visibility and usage of OA scholarly books from different publishers and on interpreting interactions related to these books. In doing so, it aims to develop methodologies and approaches to capturing and analysing these data. The outcome of this study is expected to be valuable to research funders, authors, researchers, publishers, libraries, and platforms that host, or make use of OA-monograph, all of whom are struggling to collect, analyse and interpret data that might help them to understand how OA books are disseminated, mentioned, appraised and used online. The study is also expected to help inform the development of OA monograph policies, in order to make scholarly books more accessible and enhance scholarly interaction and communication.

This study could have chosen to include a complete qualitative analysis for each book to complement the quantitative aspects of the project, and to allow for deeper exploration of some aspects of usage for the whole book set. However, a decision was made to include qualitative analysis only where there were anomalies in the data and to focus on the reasons behind these anomalies. This decision was made in order to allow for a focus on the ways in which readers are interacting with books; and to ensure that within the limited scope of a PhD project information likely to be useful for publishers, authors, libraries, platforms, and funders is not neglected. This study does not aim to examine all aspects of monographs. Rather, it is an exploration of the new insights that can be provided by quantitative data.

This study focusses on a relatively small, 28 title, set of books. Although a small set creates some limitations, it also makes it possible to dig deeper into the data relating to each of the books: combining different approaches and including different types of data sources in ways that would have been difficult with a larger set of titles and allowing for richer analysis (Moser & Korstjens, 2018). Therefore, this dissertation does not attempt to provide a a statistical study; but an exploratory pilot of the approaches and techniques capable of shedding light on the uses of OA books and a sense of what might be done using a bigger set and more resources.

1.5 Research Design

1.5.1 Theoretical Framework

This study employs an interdisciplinary approach. It draws first on citation analysis, webometrics and altmetrics approaches to understand the use and dissemination of scholarly books. There is a substantial amount of literature exploring ways to measure the visibility, usage and impact of research outputs. This literature provides helpful theoretical frameworks that inform the approaches to collecting and analysing monograph usage data developed in this thesis. This study also engages with the Crossick Report's (2015) findings by exploring how communities are interacting with OA scholarly monographs.

With the Book Citation Index, introduced in 2011 as part of the scientific citation indexing service Web of Science (WoS), and the coverage of scholarly books provided in the Google Scholar database, it has become possible to analyse the influence of monographs in academia (Harzing, 2017; Kousha & Thelwall, 2014). Moreover, webometrics methods enable the exploration of informal scholarly communication by revealing how ideas are disseminated on the web (Thelwall, 2009).

Altmetrics studies are mostly conducted on journal articles that are hosted on one platform. These studies explore mentions of research outputs on social media, including social bookmarking platforms and digital libraries, social networks, blogs, encyclopedias and other types of platforms (Torres-Salinas, Cabezas-Clavijo, & Jiménez-Contreras, 2013). Most studies in this field focus on the correlation between data obtained from these platforms and citation data collected from databases such as Scopus and WoS (Costas, Zahedi, & Wouters, 2015a; Priem et al., 2012; Yan & Gerstein, 2011).

Altmetrics is commonly used to measure attention, influence or impact. It provides valuable data on how information travels across different platforms. By combining altmetrics, webometrics, and data on their discoverability, access, mentions, citations, and ratings they receive it is possible to obtain data with a better coverage of the visibility of and interactions related to monographs across the digital sphere. Correlation analyses of these data provide valuable insight into the relationships

between these data sources. However, it is also necessary to understand the motivations behind these acts of viewing, downloading, referring, mentioning and appraising monographs.

Using citation theories (MacRoberts & MacRoberts, 1996; Merton, 1973; Nicolaisen, 2007; Small, 1978), which investigate the motives behind citation, Haustein and colleagues interpreted interactions occurring on social media. They combined citation theories with a number of social theories (Haustein, Bowman, & Costas, 2015), including social capital, attention economics, and impression management, to explain the motives behind these social media acts.

With the increase in the global usage of social media, user behaviour studies on this medium have begun to be conducted, notably in the fields of marketing and psychology. Ngai and colleagues identified 46 relevant studies on user behaviours on social media using five dominant business/management academic databases (Ngai, Tao, & Moon, 2015). Based on these papers, they developed a causal chain framework founded on the input-moderator-mediator-output model (Mohammed, Ferzandi, & Hamilton, 2010), which illustrates the causes and results of user behaviour on social media.

1.5.2 Research methods and process

This study collected and analysed data related to 28 OA monographs from the KU pilot collection. To collect data about these monographs, data sources were chosen based on data sources used in webometrics, article-level metrics (ALM), altmetrics and citation metrics. Several other sources related to monographs were also used. Data regarding discoverability, visibility, access, usage, mentions, appraisal and citations were collected and analysed. During the study, Spearman and Pearson correlation analyses were conducted to evaluate the correlations between the collected data. All data were collected and analysed using RStudio version 0.99 onward. Figures were generated using R's ggplot2 library and Tableau software.

1.5.2.1 Datasets

This study is based on data related to 28 KU pilot collection titles in the fields of anthropology, history, literature, media and communications, and politics provided by 13 scholarly publishers. The titles with their respective details, including subtitle,

author name, subject categories, and date of upload to the repository, were extracted from the OAPEN repository metafile. The datasets used in this study are shown in Table 1.1.

Table 1.1: Datasets used in this study.

Type	Detail	Date
Visibility	URLs of all web pages in which the 28 titles and author names are present	July 2017
Visibility	Links inside the web pages in which the 28 titles and author names are present	July 2017
Access	OAPEN repository access data	March 2014 – June 2017
Access	OAPEN web access statistics	March 2014 – June 2017
Access	HathiTrust web access statistics	March 2014 – June 2017
Mentions	Twitter	1 January 2014 – 1 July 2017
Mentions	Facebook	1 January 2014 – 1 July 2017
Mentions	Twitter for one UCL title	1 July 2016 – 1 April 2017
Mentions	Wikipedia	Until March 2018
Appraisal	Goodreads	Until March 2018
Usage	Mendeley	Until March 2018
Citation	Scopus	Until March 2018
Citation	Google Scholar	Until March 2018
Citation	WoS	Until March 2018

1.5.2.2 Data collection and analysis

The discoverability of OA monographs was assessed by searching for the 28 titles within the Directory of Open Access Books (DOAB) metafile, which is used in library catalogues. Subsequently, it was determined whether OA research output search engines indexed these titles. To do this, a script was used to connect to these search engines' web application programming interface (API) service which is an interface that allows programmers to access data. This script written in the R language queried the service for data related to monographs by using identifiers such as ISBN and collected the results to a file.

To evaluate the visibility of these titles, two datasets were collected. In this study, visibility is defined as the presence of the title name with its first author on web resources accessible from a search engine. The first dataset was constructed and analysed using Webometrics Analyst 2.0 software (Thelwall & Sud, 2012). Webometrics Analyst 2.0 uses the Microsoft Bing search engine through Microsoft's Azure API. A content analysis, which involved checking and categorizing each

page, was conducted on a sample of the collected web pages. In addition, a URL analysis was conducted using a script which extracted the URL parts of each page according to country, as well as the organisation to which the page belonged.

To compare the visibility of the two repositories in which the 28 titles were hosted, an R script was used to scrape links from the web pages that were identified using Webometrics Analyst 2.0. This scraped data constituted the second dataset on visibility. The number of occurrence of these repositories' addresses in this dataset was examined to evaluate their presence on the web.

The access component included three datasets from the OAPEN and HathiTrust repositories. The OAPEN repository access data consisted of download counts and web traffic statistics, whereas the HathiTrust access repository data included only web traffic statistics. Download count reports for the titles hosted on the OAPEN repository were provided by the Institutional Repository Usage Statistics UK (IRUS-UK). These reports were country-based and IP address-based reports and were downloaded and analysed using a script. Although the number of access to a document in some repositories is open to manipulation such as automatic downloads by robots, it has also a potential to uncover their relations with other data sources, such as citations (Bollen, et al., 2005; Chu & Krichel, 2007; Moed, 2005).

Using the OAPEN access data, aggregate monthly download averages were analysed. Subsequently, the average monthly downloads of each title within the same category were compared to the average monthly downloads of other English-language titles in the repository. The distributions of the 28 titles' monthly downloads were analysed and compared to those of all the English-language titles. Each title's monthly downloads were plotted and examined for patterns. The monthly download distributions for each title were analysed to identify download spikes.

The correlation between the number of domains in which each title was present (obtained in the visibility section) and each title's monthly average downloads was analysed. The correlation between the countries downloading a monograph title and the countries of the web resources featuring the monograph and its author's name was also examined. These correlation analyses were conducted in order to examine the relationship between visibility and access. A country-based download analysis

was also conducted for titles dealing with specific regions to investigate the relationship between the number of access instances from a region and the relevance of the title's content to this region.

The IP address-based reports were used to compare access instances by IP addresses associated with institutions that supported the costs of making the pilot collection titles OA to total downloads. The IP address-based reports were used to geolocate where the access originated. Subsequently, these geolocations were filtered according to states for the United States, and the correlation between number of total downloads and U.S. state population was examined.

The other datasets used were web traffic statistics from two repositories: OAPEN and HathiTrust. These statistics were collected using Google Analytics. Both repositories granted the researcher access to their usage statistics. The statistics were downloaded and analysed using a script. The OAPEN web traffic statistics were used to evaluate access to the book presentation web page of a monograph which gives information about the title as well as a book's PDF file. However, on the HathiTrust repository, the content of a monograph is provided on the web page. Therefore, the web traffic statistics for the HathiTrust repository reflect access to the monographs.

Download counts were compared to page views on the OAPEN repository to reveal the number of downloads that occurred without a user visiting the web page. A correlation analysis between unique page views of the title's presentation page and number of its download was conducted to determine whether Google Analytics data for monograph presentation pages could be used as a proxy for readers' locations. The traffic sources were analysed to identify the referring sites driving traffic to the OAPEN repository. Additionally, social network referral statistics from Google Analytics were analysed to reveal which social network platforms were most effective in driving traffic to the repository.

Unique page views on the HathiTrust website were compared to the number of downloads from the OAPEN platform. This made it possible to explore how these figures differed and to discuss the factors making OAPEN a more accessed repository than HathiTrust for the KU pilot collections.

A title was chosen as a case study, and the approaches to visibility and access discussed in the thesis were applied to this title. First, the visibility test revealed the types of sites on which the title was present, as well as these sites' countries. Subsequently, the links on three mainstream news sites where the title was present were identified. The publication dates of these news articles and the links they contained were noted. Subsequently, the spikes in downloads for this title in the OAPEN repository were recorded. Traffic sources for the presentation page during these spikes were then examined and compared with related events that occurred during the same period as these spikes. Downloads were geolocated using the IP address-based download reports. Finally, by combining these analyses with possible reasons for spikes in downloads, effective channels for the dissemination of this title were identified.

To analyse mentions, usage, appraisal and citations, data were collected from five social media platforms (Twitter, Facebook, Wikipedia, Goodreads and Mendeley) and queried on three citation databases (Scopus, WoS, and Google Scholar) in which the 28 title names were mentioned. For each platform, the number of titles covered was examined.

Using the Salesforce Radian6 platform, public mentions of the 28 titles on Facebook and Twitter were collected. The findings from Facebook were analysed according to post type for each title. The number of tweets for each title was also examined, as well as the number of Twitter users mentioning this title. 'Tweetation phases' were evaluated by analysing the number of tweets mentioning a title according to the period elapsed since the title was made OA. In addition, the Twitter capture tool (TCAT) software was used to track how Twitter users disseminated information about a title. Since TCAT software cannot access Twitter's historical data to track the connections between users in the Twittersphere, a newly published OA title from UCL Press was chosen as a case study. Nodes disseminating information about the title to other groups were visualized and identified using the network analysis and visualization software Gephi 0.9.1.

Mentions on peer-reviewed research blogs were tracked by querying the ResearchBlogging platform. Wikipedia articles mentioning the titles were manually queried using the Google search engine.

Reviews and ratings for each title on the Amazon, Google Books and Goodreads platforms were inspected through these sites' APIs using a script. It was examined whether the titles were sold via Amazon and Google Books. In addition, the format in which the books were provided on the Amazon and Google Books platforms was noted. Using its API, the Mendeley platform was examined to determine who was using these titles and how the titles were being used. To inspect how these titles were being read and to determine which titles were attracting more attention, the annotation platforms Hypothes.is and PaperHive were manually examined using each title name and also by using the PDF file of each title.

Citations of these titles were also investigated by using a script to connect to and query the citation databases Scopus, WoS, and Google Scholar. Subsequently, the yearly aggregated citations for each title on WoS were checked in order to determine whether patterns in usage could be identified. Later, a correlation analysis was conducted to check how these data sources were related to one another.

All of the data obtained through the analyses mentioned above were selected to be indicators of different interactions related to OA monographs and were classified into six categories of acts. Using the results of the correlation analyses and these classifications, the relations between these data sources were examined. This examination resulted in two kinds of acts being defined: associated acts and isolated acts.

Each data source as an indicator of a type of interaction is discussed in light of citation theories, social theories, and behaviour theories. During the discussion, the factors mediating and moderating the interactions related to the monographs are identified. Finally, the findings from this interpretation of the data sources are used to propose a revised causal chain framework that explains the interactions related to OA monographs.

1.5.2.3 Ethical considerations

The public data collected from social media contained posts and tweets from users. These data are shared publicly with the intention for them to be seen; however, no usernames are revealed in the dataset of this thesis. The only identifiable usernames relate to the Twitter data mentioning the UCL Press title. These names are included

to show how information is disseminated in the Twittersphere. There was no need to anonymize these names, since users on this platform are aware that they are being observed by strangers. Other public data collected on visibility and social media, including data obtained from platforms such as Mendeley, Goodreads, Wikipedia and citation databases, cannot be related to specific users.

In this study, access data was obtained with the cooperation of the OAPEN Foundation and Knowledge Unlatched GMBH.

The OAPEN access data processed by IRUS-UK yielded two types of reports: country-based and IP address-based reports. Although the country-based report did not include any data that could be traced to specific users, the IP address-based report contained data that could be traced to specific computers. These private data were used in this dissertation to geolocate usage, but data that can be traced to specific users are not revealed (Townsend & Wallace, 2016). The web traffic statistics collected from OAPEN did not contain any data traceable to specific users.

The Google Analytics access data provided by the OAPEN and HathiTrust foundations contained no data that could be traced to specific users or computers.

For the above reasons, no informed consent was necessary for collecting and using the data in this study.

1.6 Chapter outline

Chapter 1 briefly introduced the background of this study, the current situation of the monograph industry, metrics related to OA monographs, and the research problems. Chapter 2 reviews the literature relevant to this topic, discussing existing research relating to the influence of academic outputs, focussing on citation theories, webometrics and altmetrics methods. This chapter also describes the current situation in the OA monograph market and identifies the gaps in research relating to its usage.

Chapter 3 discusses methods for capturing the discoverability, visibility and usage of OA monographs in a digital landscape. Monograph usage data collected from various repositories using various techniques is examined and analysed. Monographs' visibility on the web is captured using webometrics approaches. Finally, the chapter considers the usefulness of these statistics in identifying readers'

locations, the ways in which monographs are discovered, and the channels that are effective for disseminating information about these books.

In Chapter 4, a deeper analysis of the social media mentions of these titles is conducted using an altmetrics approach. Social network mentions, Wikipedia page citations, ratings from book rating sites, usages from bookmarking services and annotation services, and citations are collected and analysed. Subsequently, a correlation analysis between all the data sources collected in Chapters 3 and 4 is conducted in order to identify the relationships between them. Finally, the interpretation of these metrics is discussed, along with the challenges and issues encountered during their collection and analysis.

In Chapter 5, each data sources used in Chapter 3 and 4 are classified according to the interactions they are indicator of. Subsequently, these indicators are classified into the types of act that have been proposed. These acts are then interpreted according to citation and social media theories. This chapter introduces additional social theories that are relevant to this study, in addition to the theories that are discussed in Chapter 2. A causal chain framework is proposed to help us to expose the factors that lead to these acts occurring in relation to OA monographs.

Chapter 6 discusses the issues facing OA monograph publishers, focussing on the discovery, visibility and consumption of monographs. This chapter highlights various issues and challenges relating to monographs, including standards for identifiers, interpretation of data sources, and data ethics. It then discusses the potential usefulness of this dissertation for various parties. Lastly, the chapter identifies future work that should be conducted on this subject.

2 Literature Review

This chapter reviews existing studies on the influence of academic research outputs. The need for filters and a way to identify the impact or influence of research outputs has created new ways to collect and interpret different kind of data related to them. At first, these data were citation-based metrics. Subsequently, with the increased usage of the web and social media, new approaches were created, including webometrics and altmetrics.

These approaches are valuable in order to capture the footprints of academic research outputs in conventional scholarly literature and also on social media. They also provide an indication of how scholarly literature is disseminated on social media and how people are interacting with it. In addition to webometrics and altmetrics approaches, citation theories approaches are useful for interpreting behaviours on social media. Combined, these approaches provide a more detailed picture of OA monographs, which until now have received little attention compared to journal articles. Lastly, this chapter explores the current situation of monographs, and in particular OA monographs, to shed light on why OA monographs should be approached differently to OA journal articles.

2.1 Citations

Faced with an overload of information, scholars have long relied upon filters to help them identify work that is relevant, trustworthy and useful for specific research purposes. These filters were at first manually compiled compendia and corpora. Over time, however, the volume of published scholarly materials increased, and problems arose with alphabetically classified indexes, for example, differences between the subject approach of a document's author and the subject approach of the researcher seeking information. In response to these issues, Garfield proposed a citation index, which would aggregate references in scientific articles and be used as a retrieval tool for scientific information (Garfield, 1955; Priem et al., 2012). Weinstock stated that this would solve the semantic problems associated with traditional subject indexes by using citation symbology rather than words to describe

the content of a document (cited in Noruzi, 2005). Later, Garfield suggested using this citation-based metrics approach to indicate the degree of influence or impact of ideas from specific publications (Garfield, 1964).

In the 1970s, following the increase in the use of citation metrics in the academic world, information scientists, sociologists and others argued that there was a need for a citation theory, which could explain why and how authors cite. They also began investigating the symbolic characteristics of citations, by examining author's interpretation of a cited work (Nicolaisen, 2007).

Citations represent different things to different researchers (Cronin, 2016). For example, citations have been described as scholarly bricks (Solla Price, 1986), as signposts left behind (L. C. Smith, 1981), as applause (Nelson, 1997), as gifts (Hagstrom, 1982), as forms of reward or income (Ravetz, 1973), as tools of persuasion (Gilbert, 1977), as traces of conversations between texts (Czarniawska, 1997), as pellets of peer recognition (Merton, 2000) and as frozen footprints on the landscape of scholarly achievement (Cronin, 1981). According to Cronin, this is due to citations' chameleon nature, which allows them to indicate different things according to the context in which they are used. Based on Czarniawska-Joerges' idea (1997) Cronin states that citations can be seen as a conversation between texts (Cronin, 2016). Thus, to capture these conversations across academia, such as who talks to whom, which conversations matter most, and how a topic, theory or a researcher's thesis evolves over time, Cronin suggests tracing relationships amongst academic citations and visualizing the networked threads.

Citation metrics attempt to quantify citations of scholarly materials, making it possible to draw conclusions about the use of individual articles, impact of authors, and academic journal titles for scholars, institutions, research funders, organizations and other parties. Zunde (1971) defines the three main applications of citation analysis as:

1. Qualitative and quantitative evaluation of scientists, publications, and scientific institutions
2. Modelling of the historical development of science and technology
3. Information search and retrieval

2.2 Citation theories

Beginning in the 1960s, information scientists and other researchers began questioning why and how authors cite in their research outputs. In order to shed light on the motives for citing, a number of theories have been proposed. In the following sections, three citation theories are discussed. These are the normative theory, the social constructivist theory and the concept symbol theory.

2.2.1 The normative theory

According to the normative theory, citing behavior is guided by the four basic norms upon which the sciences are supposed to be founded (Merton, 1973). These four basic norms are universalism, communism, disinterestedness and organized skepticism. According to this theory, when assessing the work of other researchers, scientists are not influenced by the author's personal or social attributes, such as sex, race, nationality, religion, and class (universalism). The normative theory also states that scholars cite materials to acknowledge the value of a colleague's work to their own research (communism). According to Merton, "communism" refers to the sense of common ownership of goods. He supposes that intellectual property rights in science are usually minimal due to the scientific ethos, so the scientist's only claim on their intellectual property is recognition and esteem. These norms also state that when citing, scientists are not looking for personal gain (disinterestedness). Lastly, the normative theory proposes that scientists should approach their own work with the same skepticism that characterizes their approach to the work of other scientists (organized skepticism).

In fact, science has not always been governed by the norms Merton identifies. Small states that these norms have evolved from exemplars of good practice or as reactions to new social realities (Small, 2016). For example, Small argues that the norm of openness is probably more important at present, since in the 17th century scientists kept their discoveries secret in order to receive proper credit. He also asserts that the invention of the scientific journal in the 1600s may have developed the norm of communism. Moreover, Small points out that the norm of disinterestedness evolved as a result of punishments for scientific fraud, which can call into question a scientist's trustworthiness and consequently endanger their career.

2.2.2 The social constructivist theory

An opposing area to the normative theory is the social constructivist theory. The social constructivist theory argues that scientists cite other works for reasons that have nothing to do with the intellectual debt referred to in the normative theory. For social constructivists, a closure in a scientific dispute is the outcome of an agreement where one side convinces the other side by means of persuasion. Thus, the outcome is not entirely independent of personal and social factors. (MacRoberts & MacRoberts, 1996; Nicolaisen, 2007). Based on the research conducted by Haustein, Bowman, and Costas (2015), this study uses four main sources of social constructivist theories which deviate from normative theories: the persuasion hypothesis, perfunctory citations, the Matthew effect, and negational citations. According to the persuasion hypothesis, researchers consider citations a tool for persuading the scientific community of the value of their work, which can be achieved, for example, by citing established authorities to gain credibility. However, Henry Small contradicted this argument by pointing out that Watson and Crick's 1953 paper on the structure of DNA has only six references, Darwin's *Origin of the Species* has even fewer, and Einstein's 1905 paper on special relativity contains no references at all. Small argues that carefully selected references are not enough to make a trivial paper convincing (Small, 2016).

Perfunctory citations are nonessential, superficial, redundant or even incorrect citations, such as references that merely contribute to the chronological context of the citing paper. Another deviation from the normative theory is the Matthew effect, where scientists with wider recognition find it easier to gain more recognition. The social constructivist theory also recognizes negational citations, where an author cites a paper to challenge or contradict the work.

However, there are also criticisms of the social constructivist theory. Small states that instances in which authors suffer psychological discomfort as a result of citation missteps are evidence that the Mertonian norms are adhered to. One example is an author's embarrassment resulting from the failure to cite an obvious precursor. If the author's colleagues become aware of this, the author risks having psychological or social sanctions imposed upon them. Small also argues that in a norm-governed publication world, instances in which a prior author's work is misquoted or distorted

and which can be classified as “constructivist” are relatively rare (Small, 2004; Small, 2016). Furthermore, referencing appears to fit the model of strong reciprocity, where generous citation is rewarded and non-citers are sanctioned.

According to Small (2016), cooperation and competition are pervasive in science. Thus, science incorporates both selfish and altruistic individuals, but each individual may also incorporate both of these tendencies. For this reason, this study uses both normative theory and social constructivist theory to interpret monograph-related metrics and to understand the factors affecting mentioning behaviors.

2.2.3 Semiotics of citation

To better understand the role of citation, Cronin suggests the use of semiotics (Cronin, 2000). Semiotics can be summarized as the systematic scholarly analysis of sign systems (Wouters, 2016). Charles Sanders Peirce, whose semiotics approach Cronin employs, wrote the following: “Consider the practical effects of the objects of your conception. Then, your conception of those effects is the whole of your conception of the object”. So, our conception in fact is not the actual object of reference but meanings. Cronin using this approach on the link between signs and real world object tried to expose the link between citation and research behavior. (Wouters, 2016).

Before proceeding, it is useful to clarify the distinction between reference and citation. If a paper R contains a bibliographic note using and describing paper C, then R contains a reference to C and C has a citation from R (Solla Price, 1970). Thus, reference is a backward-looking concept, while citation is a forward-looking one (Egghe & Rousseau, 1990).

Wouters stresses the distinction between these two terms by stating that they are different signs (Wouters, 1999). According to him, they are positioned as different objects. He states that a citation emerges in an act of “semiosis” (the creation of a novel sign) from the reference by its registration in a citation database. Thus, the creator of a citation is not the researcher, but the producer of a citation index.

To better understand these distinctions, Cronin (2000) uses Peirce’s “sign triad”. The triad consists of three dimensions. These are: (i) the carrier of the meaning (sign-

vehicle); (ii) the meaning or concept referred to (interpretant); and (iii) the object pointed to (referent).

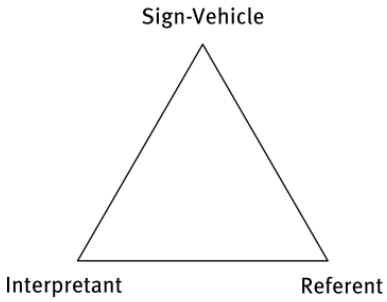

Figure 2.1: The sign triad. Reprinted from "Semiotics and citations" by P. Wouters, 2016. In C. Sugimoto (Ed.), Theories of informetrics and scholarly communication, p. 75. Walter de Gruyter GmbH & Co KG.

Cronin applied this triad to bibliographic reference and citation. In this case the sign vehicle is the embedded reference which is the carrier of the meaning. The situated meaning or the concept is addressed by the reference that can be located in a specific part in the cited text. Finally the referent can be either the full reference in the bibliography, or the cited text (Wouters, 1999). The triads are as follows:

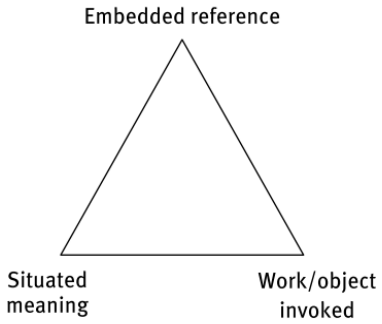

Figure 2.2: Bibliographic reference sign triad. Reprinted from "Semiotics and citations" by P. Wouters, 2016. In C. Sugimoto (Ed.), Theories of informetrics and scholarly communication, p. 75. Walter de Gruyter GmbH & Co KG.

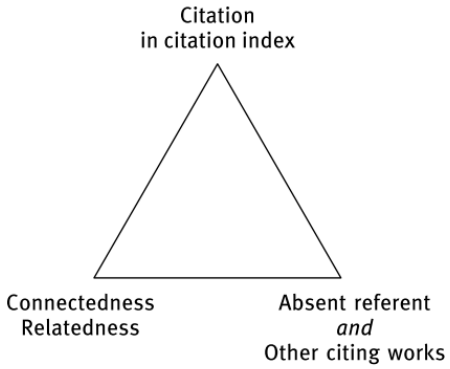

Figure 2.3: Citation sign triad. Reprinted from "Semiotics and citations" by P. Wouters, 2016. In C. Sugimoto (Ed.), Theories of informetrics and scholarly communication, p. 76. Walter de Gruyter GmbH & Co KG.

Since scientometricians usually neglect to acknowledge the different configurations of the sign they are discussing, different realities have been collapsed into one another (Wouters, 2016). For example, the lower left corner for the embedded reference (Figure 2.2) may be "situated meaning", and therefore can explain why the author made the reference in the first place. However, citation analysis does not deal with embedded references, but with aggregated citations. In this case, in the first step, the embedded reference is transformed into the reference as listed in the bibliography. Subsequently, in the second step, this list is inverted in the production of the citation index. Since the reference is decontextualized twice in this two-step procedure, the situated meaning cannot be aggregated and as a result the meaning is lost. This is why Wouters argues that applications of citation analysis can no longer be based on the original situated meaning of the embedded reference.

Thus, it does not make sense to justify citation counts in an evaluative context by claiming that the number of citations expresses the scientific community's opinion about the value of a particular work or author. A citation's meaning cannot be extracted from the citation itself, because the citation object as number is decontextualized, and, in Wouters' words, is an "underdefined proto-object".

In summary, Wouters argues for the adoption of "material semiotics" in informetric research, which explores the quantitative aspect of information. In material semiotics, relations are mapped as material (between things) and at the same time as semiotic (between concepts). In this dissertation, the reference, citation, and "citation as part of the citation index" are seen as ontologically different but related objects.

2.2.3.1 The concept symbol theory

Another semiotic approach to citation is the concept symbol theory, which was pioneered by Henry Small (Sugimoto & Larivière, 2018). This theory approaches "cited documents as concept symbol" (Small, 1978). Small states that the citing is a symbolic act in which authors associate particular ideas with particular documents. Small notes that when citing a document, the meaning of the cited document is limited to a few sentences, which would result in a distortion or oversimplification of ideas. His theory helps to clarify the meaning of and motivations underlying citations. Small's approach can be used for the construction of indicators, which are linked to a concept and represent a valid measurement of that concept.

To quantify a concept, we need to create indicators that represent a valid measurement of the concept. This is because the indicator is not the concept itself, but a proxy that serves as a way of measuring how the reality behind the concept changes over time and/or across space. For the representation to be successful, the relationship between the indicator and its corresponding concept should be strong. This allows unobservable variables (e.g., a monograph's visibility or impact) to be quantified in a statistically valid way, by using indicators or measurable variables (e.g., the number of pages mentioning a monograph or the number of citations referring to the monograph) (Lazarsfeld, 1958).

According to the science historian Gingras (2014), the indicator should also match the rate of change of the concept being measured. He gives the example of an indicator ranking an institution that moves in a single year from, say, 12th to 18th or 12th to 9th. This would strongly suggest that the indicator representing this movement is defective, because the quality of an institution is unlikely to rise or fall significantly during the course of a single year. For this reason, an indicator should be chosen so that it is sensitive to the intrinsic inertia of the object measured.

Gingras also stresses that some indicators lack homogeneity, such that different heterogeneous indicators are incorporated in one indicator (Gingras, 2014). For example, the Altmetric Attention Score, which aims to measure the attention paid to a research output, combines various values collected from social media platforms, including Mendeley, Twitter, Facebook and newspapers, making the score uninterpretable (Sugimoto & Larivière, 2018).

Cronin asserts that the idea of unifying theory of citation is nonsensical and instead argues for "a number of partly contradictory, and partly overlapping sets of citation theories, each emerging in a particular set of knowledge practices" (Cronin, 2000). Therefore, as stated above, this dissertation adopts various partly contradictory, partly overlapping theories to analyze data collected from different sources.

In addition to the issues mentioned above, there are limitations associated with citation indexes when it comes to fields outside of the hard sciences and non-journal publications. In particular, citation indexes have been criticized for their poor coverage of monographs and monograph chapters, especially for titles published in the HSS. Moreover, in contrast to journal or proceedings articles, authors cite book material in different ways, which require different treatments when analyzing citations (Purnell and Glänzel, 2013).

2.2.4 Impact vs. quality of a research output

In recent years, evaluations of research activity have become an increasingly common consideration in decisions regarding research funding and in assessments of researchers according to performance criteria. This form of evaluation is intended to stimulate research productivity. Evaluations of research outcomes usually involve two proxies: a publication's citations and the impact factor of the journal in which it was published. However, the use of citations as indicators of quality or impact is controversial (Leydesdorff, Bornmann, Comins, & Milojevi, 2016). As Garfield (1979) says,

> People talk about citation counts being a measure of the "importance" or "impact" of scientific work, but those who are knowledgeable about the subject use these words in a very pragmatic sense: what they really are talking about is utility. A highly cited work is one that has been found useful by a relatively large number of people, or in a relatively large number of experiments... Conversely the citation count of a particular piece of scientific work does not necessarily say anything about its elegance or its relative importance to the advancement of science or society.

Thus, it is possible to say that a paper's citations do not reflect its quality or the nature of the work, and also say nothing about its utility or impact. In addition to

these points, MacRoberts and MacRoberts argue that citation behavior is prone to errors and biases of various kinds. For this reason, they assert that citation analysis is an illegitimate tool for research performance assessment (MacRoberts & MacRoberts, 1996). Wouters suggests that semiotic approaches are likely to provide a better framework for understanding the historical development of evaluative bibliometrics and the pervasive influence of the citation index on how quality is defined in the scientific and scholarly system (Wouters, 2016).

2.3 The need for new types of filters

With the introduction of digital technologies and the spread of the Internet, intra- and inter-document searchability and navigation have improved (Kling & Callahan, 2003), leading to considerable growth in the volume and diversity of academic literature in all disciplines. In addition, these developments have created more open, transparent, and diverse possibilities for scholarly communication, and has led to journal and monograph publishers' adoption of OA and to researchers sharing their datasets online. Moreover, new platforms have provided tools and environments that enable researchers to collaborate online with their peers. As a result, the ways in which research communities interact with one another, as well as with one another's work, are becoming more digitally visible. Discussions between researchers can now be viewed on blogs and across social networking sites, including Twitter, Facebook, and dedicated academic social media platforms like academia.edu.

These developments have created a growing information overload. Researchers are increasingly struggling to navigate the noisy scholarly communications landscape and have begun to question the value and effectiveness of filters like citation-based metrics. Citation analysis and bibliometric methods more generally are increasingly being criticized for oversimplifying the concept of scientific research productivity and quality, as well as encouraging gaming behaviors such as 'salami publishing' amd honorary authorships and allowing for the misuse of indicators (Haustein, Sugimoto, et al., 2015). Moreover, established filters such as journal impact factors and citation metrics are of limited value in this context, particularly when it comes to understanding the ways in which digital publications are being used across globally distributed research communities (Priem, 2014). These developments have created a demand for new filters to guide scholars and help them engage with content that is

useful to them. In an effort to respond to these demands, multiple filters have been proposed, including webometrics and altmetrics methods (Priem, Groth, & Taroborelli, 2012).

2.3.1 Webometrics

In scientometrics, which is the quantitative analysis of science, researchers analyze information about sets of scientific documents extracted from a publication database or citation index. Following the introduction of the Internet, researchers attempted to apply scientometric methods to the web. In 1997, Almind and Ingwersen (1997) coined the term "webometrics", which is an information science concerned with the analysis of the quantitative aspects of the web. Thelwall (2009) defines webometrics as the measurement of different aspects of the web, including websites, web pages and parts thereof, words in web pages, hyperlinks and web search engine results. He argues that, as a general rule, qualitative and quantitative webometrics techniques should be used together. Using qualitative techniques alone may risk overlooking the bigger picture, due to the necessarily small-scale nature of these techniques. On the other hand, using only quantitative techniques risks a superficial or misleading analysis if these approaches are not complemented by a supporting qualitative analysis. Thelwall points out that a content analysis component is especially important to help interpret quantitative data if more in-depth qualitative methods are not used.

Webometrics attempts to shed light on the impact or spread of ideas on the web. By using search engines to count the number of web pages or blogs that mention a key phrase, large-scale coverage is possible, and the data collection process is passive and relatively cheap. On the other hand, there are also some disadvantages of webometrics, such as the lack of control over the web "sample" used and the restriction to web-only data. In summary, Thelwall argues that the web can often be used to obtain quick, indicative results, and can in some cases be used for more in-depth studies as well.

Thelwall also states that webometrics provides the possibility to explore informal scholarly communication which could help investigate the research process. Scientometricians have often investigated the development and mechanism of

science using statistical mathematical methods with the aim of understanding processes or change and improve infrastructure. Thelwall argues that the processes involved in research, which include informal discussions between investigators, talks at conferences, and the formation of research teams, typically go unpublished. In fact, these research processes were studied before the introduction of the web, by sociologists of science and information scientists. However, these investigations were time-consuming and limited in scale. The difficulty of gathering sufficient data made scale a challenge, because these pre-web studies were conducted by observing researchers at work or interviewing conference attendees. The advent of the web has made it possible for investigations of scientific knowledge production processes and communities to be conducted at a much larger scale, because a significant proportion of informal scholarly communication now occurs online or at least leaves an online trace. For example, conference proceedings are often posted on the web, research groups tend to have informative websites, and academic debates can occur on blogs or discussion lists with web-based archives.

The ability to examine web links using the AltaVista search engine, which was the most-used search engine prior to Google, was influential because hyperlinks are similar to academic citations in structure. Academic citations point from a source document to a target document, just as web links point from a source page or document to a target page or document. The similarities between hyperlinks and citations, together with universities' early adoption of the web, resulted in the emergence of a number of research goals. Researchers attempted to assess whether hyperlinks could be used in similar ways to academic citations and also attempted to determine the validity of using link counts derived from AltaVista data for specific ideas just like for counting citations for articles (Thelwall, 2001).

According to Thelwall, "cybermetrics" is a term used to describe research that is essentially the same as that conducted in webometrics. *Cybermetrics* was the name of an electronic journal launched in 1997. This term was particularly popular in Spain, where, according to Thelwall, the word "webometrics" has the unflattering popular connotation of "egg/testicle measurer". The difference between the two terms was resolved by allowing cybermetrics to be more general. As such,

"cybermetrics" refers to non-web Internet research, such as email or newsgroup studies, in addition to web research (Björneborn & Ingwersen, 2004).

Long after its coinage, the term "webometrics" was given its accepted definition of "the study of web-based phenomena using quantitative techniques and drawing upon informetric methods" (Björneborn and Ingwersen, 2004). According to Thelwall (2009), the importance of this definition was its inclusion of informetrics methods as the main characteristic of webometrics. Informetrics is a term used in information science to refer to quantitative research centered on measuring information, such as citation analysis. The relationships between different metrics studies are shown in Figure 2.4 (Björneborn & Ingwersen, 2004).

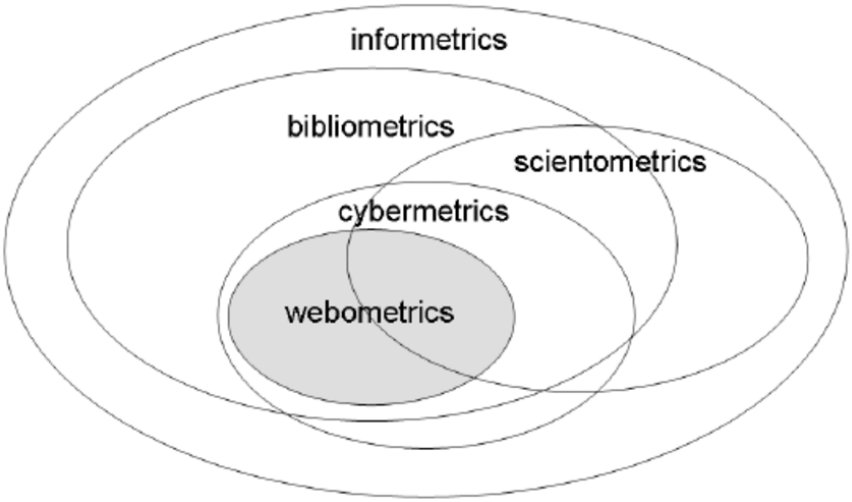

Figure 2.4: The relationships between different metric studies. Reprinted from "Toward a basic framework for webometrics" by L. Björneborn and P. Ingwersen, 2004, Journal of the American society for information science and technology, 55(14), 1216–1227. Copyright 2004 by John Wiley and Sons.

2.3.2 Webometrics research

Webometrics techniques include impact assessment, link analysis, and blog searching. Web impact assessment, which includes analyses of web mentions and content analysis, aims to evaluate the impact of documents or ideas on the web. Link analysis, on the other hand, focuses on the hyperlinks between web pages. Blog searching is an investigative technique that involves querying keywords on blog-specific search engines.

2.3.2.1 Web impact assessment

Web impact assessment, in a webometrics approach, involves evaluating the web impact of documents or ideas by counting how often they are mentioned online. The general idea is that, all other factors being equal, documents or ideas with more impact are likely to be mentioned more online. This concept originates from a study conducted in 1998, which counted how often prominent academics were mentioned online and also examined the contexts in which they were mentioned (Cronin, Snyder, Rosenbaum, Martinson, & Callahan, 1998). Vaughan and Shaw (2003) found that the web impacts of journal articles in the field of library and information science were significantly correlated with the bibliographic database service Institute for Scientific Information (ISI) citations (now WoS), and just under half of these citations seemed to directly reflect scholarly or educational impact.

Thelwall suggests that it is possible to investigate similar books' web impacts to compare their influence or spread. He argues that examining sales figures in a particular territory is not an efficient strategy for understanding the context in which books are read; nor are purchasers' evaluations of a book's quality and usefulness. Measures of how often a book is mentioned in blog posts and online reviews and indications of the countries from which posts and reviews originate can provide useful information about a book's reception and use. Thelwall also emphasizes the importance of qualitative investigation and content analysis in attempts to interpret the significance of web mention data. This is particularly important in the context of negative reasons for web page creation, such as spam marketing. The end result of this content analysis will be a set of categories and an estimate of the number of search results that fit each category. These categories can be national origins or industrial sector. The categorizations should address who created the citing pages, as well as the purpose of the citing pages. When the content analysis is complete, it should be used to complete sentences like "Document/idea X was mainly mentioned online by A and B" or "Document/idea X was mainly mentioned online because of A and B" (Thelwall, 2009).

2.3.2.2 Link analysis

There are two main types of webometrics link analysis: link impact assessment and link relationship mapping. Link impact assessment involves counting the number of

links to a web resource, such as a web page, a document, an information source from a database or a web service. The number of links can be counted by using a search engine and looking for website addresses or in technical terms their uniform resource locators (URLs) outside of the target domain name. To interpret the data from a link impact assessment for a collection of websites, Thelwall suggests a list of analyses that can be done. One is to compare overall hit counts to determine the pages, sites or countries that attract the most links and the reasons for this. Another is to compare the type of site, such as blog, news portal or academic site, which links most frequently to each page or site and commenting on why this pattern is observed. It should be remembered that the counts from search engine results do not report on the entire web, but only the part of the web that the search engine covers. In addition, the search engine may not reveal all the information it has gathered.

Another type of link analysis is link relationship mapping, which is used to illustrate the pattern of links within a collection of websites. Link relationship mapping produces a network diagram, with nodes representing websites and arrows between nodes representing the links between them.

2.3.2.3 Blog Searching

Blogs are a valuable source for webometrics research, because they contain time-stamped postings. Apart from blog-specific search engines, some science-only aggregators also exist, such as the ResearchBlogging.org (Shema, Bar-Ilan, & Thelwall, 2014) platform. These platforms aggregate blog posts that refer specifically to peer-reviewed research and these blog posts can be queried on these platforms. Therefore, these platforms provide tools for identifying interests and trends among the general public and within academia.

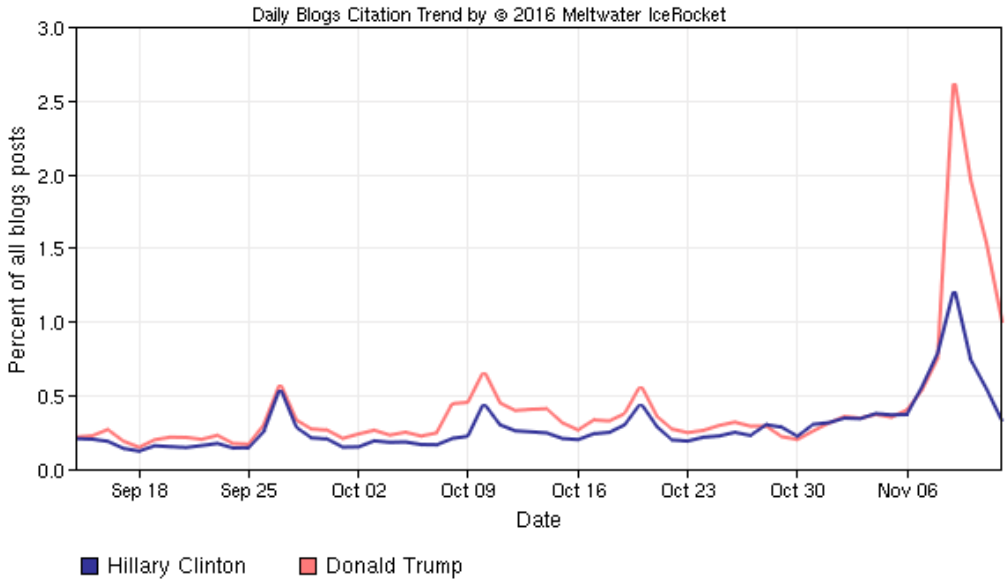

Figure 2.5: Blog trend graph from Meltwater Icerocket search engine.

Blog trend graphs, produced as in Figure 2.5, are useful for a number of purposes, including:

- Identifying the starting point for discussion of an issue
- Identifying key events related to a broader issue
- Identifying long-term trends
- Performing comparative time series analysis

Blog searching can be useful for identifying fluctuations in public interest in a topic and can also reveal the causes of these fluctuations. A blog searching analysis can be supported by a Google trend search to confirm that the trends identified are not particular to blogs.

2.4 Altmetrics

In an effort to draw attention to the need for filters aside from webometrics capable of encompassing online activities, Priem and others coined the term “altmetrics”, short for alternative metrics (Priem et al., 2010). Altmetrics is a sub-discipline of scientometrics, that is, the science of measuring and analysing science and scientific research (Fenner, 2014). Although there is no widely accepted definition of the term, “altmetrics” is generally used to refer to metrics that are concerned with the influence of a subject on social media, using indicators of visibility and awareness such as mentions (Galligan & Dyas-Correia, 2013; Holmberg, 2014). In fact, in

2000, Blaise Cronin already predicted new approaches to capture new forms of signals to be used as indicators of social influence (Cronin, 2000). In Cronin's own words,

> The web is giving rise to new modes of communication, representation, recommendation and invocation. The ways in which, and reasons why, individual researchers and scholars are mentioned, or linked to on the web, are multifaceted. It is conceivable that novel forms of signaling will evolve, which could also be used as indicators of cognitive or social influence within specific disciplines or communities of professional practice.

According to Moed (2015), three drivers led to the development of altmetrics. The first driver was the increasing awareness of the multidimensionality of research performance. Organizations started to look for alternative and broader ways of measuring the productivity and performance of individual researchers (Bar-Ilan, 2014). The second driver was the change brought about by the development of information and communication technologies (ICT). The final driver is the open science movement, which is defined in the UNESCO portal as

> The movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional. It encompasses practices such as publishing open research, campaigning for open access, encouraging scientists to practice open notebook science, and generally making it easier to publish and communicate scientific knowledge. ("Open Science Movement")

According to Bornmann (2014b), since the 1990s, the trend in science policy has been to no longer assume that society is benefiting from science pursued at a high level. It is now expected for the benefits of scientific research to society to be demonstrated. As a result, funding organizations such as the U.S. National Science Foundation require evidence that the projects they support are beneficial to society and have an impact on it. Bornmann and Haunschild argue that broadening the notion of impact from citations to societal effects represents a scientific revolution in scientometrics (Bornmann & Haunschild, 2017). Altmetrics is a good candidate to be used as a societal impact indicator.

Due to the increased adoption of altmetrics in organizations, the National Information Standards Organization (NISO)—a non-profit organization that maintains and publishes technical standards related to publishing, bibliographic, and library applications—published a report in 2014 on best practices in altmetrics (NISO, 2014). This document is intended to help organizations that wish to use altmetrics to ensure the consistent generation and aggregation of altmetrics data across the community.

2.4.1 Altmetrics and impact

In addition to tracking engagement with websites, blog posts, software, and data sets by using mention, download or tweet statistics, altmetrics approaches are used to track the impact of articles and books (Piwowar, 2012). However, how the term "impact" should be defined in relation to scholarly works is controversial.

In the *Metric Tide* report (Wilsdon et al., 2015), the Higher Education Funding Council for England (HEFCE) defines impact outside of academia as "an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia". This definition has been criticized as being too narrow. Instead, Tinkler (2015) suggests that we use other metrics to explore the richness and diversity of the outcomes of research, recognizing that impact is a multi-dimensional concept, which changes over time and differs across disciplines and sectors. Therefore, due to its wide range of data sources coverage that will be discussed in the next section, altmetrics can be a good proxy candidate for the impact of and engagement with research outcomes.

Galena and Dyas-Correia point out that altmetrics offers the possibility of obtaining insights into impact that could not be obtained before (Galligan & Dyas-Correia, 2013). According to these authors, the issue here is to find an appropriate fit for altmetrics, so that they can fulfill their potential in specific circumstances. In addition, they stress that altmetrics users themselves must determine how to employ particular altmetrics in order to deliver the correct context.

According to Sugimoto (2015), the term "impact" is often misappropriated by the altmetrics community, because this term connotes broader engagement and a more transformative effect than is currently obtainable using altmetric data. For example,

tweets or Mendeley saves do not indicate that a monograph has had a strong effect on the user. Thus, Sugimoto suggests that rather than being understood as a proxy for impact, altmetrics should be used to complement existing metrics and therefore to expand the tools available to provide insight into the dissemination of science. Therefore, with knowledge of the strengths and limitations of the available tools and data, it is possible to construct richer narratives of the ways in which scholarship is diffused and the impact it has on society (Konkiel, Sugimoto, & Williams, 2016).

Therefore, whether altmetrics measure "impact" or not, they offer new insights into the usage and diffusion of scholarship. This study uses altmetrics in order to understand interactions with monographs and to expose how monographs are diffused across the digital landscape.

2.4.2 Altmetrics data and their categorization

Researchers and institutions categorize altmetrics differently. Torres-Salinas, Cabezas-Clavijo, and Jimenez-Contreras (2013) categorize altmetrics data as follows: (1) social bookmarking and digital libraries (e.g., Mendeley), (2) mentions on social networks (e.g., Twitter and ResearchGate), (3) mentions on blogs (e.g., Wordpress and Nature Blogs), (4) mentions in encyclopedias (e.g., Wikipedia), and (5) mentions in new promotion systems (e.g., Faculty of 1000). However, ImpactStory (Piwowar, 2012) classifies altmetrics as scholarly and public metrics, as shown in Table 2.1.

Table 2.1: ImpactStory classification of Altmetrics.

Viewed	Saved	Discussed	Recommended	Cited
PLOS HTML	CiteULike	Nature Blogs	F1000 Prime	Scopus
PLOS PDF	Mendeley	ScienceSeeker		WoS
PLOS XML		ResearchBlogging		CrossRef
PMC HTML		Wikipedia		
PMC PDF		Twitter		
		Facebook		

The Public Library of Science (PLOS), which started to use article-level metrics (ALM) for all of their journal articles in 2009, classifies their metrics into views, saves, discussions, recommendations and citations, as shown in Table 2.2 (Lin &

Fenner, 2013). Although the terms “altmetrics” and “ALM” are usually used interchangeably, they differ from each other in two important ways: (1) in contrast to altmetrics, ALM include citation and usage data; and (2) altmetrics can also be applied to other scientific materials, such as research data (Fenner, 2014).

Table 2.2: PLOS Article-Level Metrics classifications.

	Scholars	Public
Recommended	Citations by editorials, F1000	Press article
Cited	Citations, full-text mentions	Wikipedia mentions
Saved	CiteULike, Mendeley	Delicious
Discussed	Science blogs, journal comments	Blogs, Twitter, Facebook
Viewed	PDF downloads	HTML downloads

In 2014, Altmetric.com defined four types of altmetrics data sources. These are: (1) social media (e.g., Twitter and Facebook); (2) reference managers or reader libraries (e.g., Mendeley or ResearchGate covering scholarly activity); (3) various forms of scholarly blogs (Shema et al., 2014) reflecting scholarly commentary; and (4) mass media coverage, for instance, daily newspapers or news broadcasting services that inform the general public (Moed, 2015). In 2018, they added the following new types of data sources (“What outputs and sources does Altmetric track?,” 2018): (5) policy documents (governmental and non-governmental); (6) post-publication peer-review forums (e.g., PubPeer, Publons); (7) patent citations; and (8) other online sources (e.g., YouTube, citations).

Priem (2014) classifies altmetrics sources by type and audience, which can be general users or scholarly users, as shown in Table 2.3.

Table 2.3: Altmetrics sources by type and audience according to Priem (2014).

Type	General Users	Scholarly Users
Recommendation	Web-based mainstream media	F1000
Citation	Wikipedia	Citation in peer-reviewed literature
Conversation	Twitter, Facebook, blogs	Scholarly blogs, article comments, tweets from scholars
Reference	Social bookmarking	Social reference managers
Reading	HTML views	PDF downloads

Moed (2015) states that there should be some distinction between audience, such as between scholarly and societal impact, and type, such as between peer-reviewed and non-peer-reviewed manuscripts. He also argues that there is a difference between downloads and citations. Moed stresses that a download of the full text of a document does not indicate that it has been read. Although both downloads and citations may influence one another in multiple ways, they are in fact distinct concepts. He also points out that the user (reader) and the author (citer) population may not coincide. He suggests that comparisons of citation counts and full-text downloads of research articles may provide more insight into citation practices and usage behavior.

2.4.3 Evaluative metrics

Generally, it is not particularly informative to report only the number of citations or the average citation frequency of a particular unit, such as a researcher or institution. The number of citations should be related to some reference group so that statements regarding the unit's citation frequency are made in relative terms. Furthermore, the publications in the reference group must be within the same subject area and of the same publication type, and must have been published in the same year. According to Schubert and Braun (1996), there are three primary methods used to define such reference groups. These are: (1) journal normalisation, which includes in the reference group publications in the journals that the analyzed unit publishes in; (2) field normalisation, which includes in the reference group publications in the subject fields that the analyzed unit publishes in; and (3) ad hoc normalisation, where the reference group is created according to a particular criterion, such as bibliographic coupling, where publications related to each other by mutually shared references are selected.

However, Moed and Halevi (2015) underline that indicators that are appropriate in one context may be invalid or useless in another. For example, they state that publication counts are useful tools to identify the active researchers among staff members, but are of little value for comparing the research performance of these active researchers.

Because research evaluations are relying increasingly on metrics, which are often incorrectly applied, the *Leiden Manifesto*, which proposes ten principles for guidance regarding research evaluation, was published in 2015 (Hicks, Wouters, Waltman, De Rijcke, & Rafols, 2015). In this manifesto, the authors warn that there is variation across fields in publication and citation practices. They suggest that a good approach would be to select a suite of possible indicators and allow fields to choose from among them. Regarding indicators, they give the example of social scientists requiring books and literature written in their national language to be included in their publications, whereas computer scientists require conference papers to be counted. They also state that citation rates vary by field as well. For example, the top-ranked journals in cell biology have impact factors of around 30, whereas top-ranked journals in mathematics have impact factors of only around three.

In addition to selecting suitable indicators, benchmarking an OA monograph requires allowing the monograph's engagement rate to vary according to when the monograph was made OA, the language in which it is published, its subject, and other attributes. For example, a monograph made OA at the beginning of the 2000s will have a different engagement pattern compared to a monograph made OA in 2018. This is because, in the early 2000s, some data sources used in altmetrics had only just begun operating, if they had started at all. These sources include Facebook (started in 2004), Twitter (started in 2006), and Mendeley (started in 2008). The rate of engagement on social media platforms is also increasing annually. Therefore, when performing a comparison in a specific context, it is important to choose appropriate indicators and a suitable reference group.

2.4.4 Benefits and Limitations of Altmetrics

Wouters and Costas (2012) identify four benefits of altmetrics measurement: diversity, speed, openness and broadness. According to these authors, altmetrics is not limited to scholarly articles, but can also be used for other types of research materials (Bornmann, 2014a). Unlike citations, altmetrics draw on different sources of data, and therefore make it possible to interrogate different forms of impact. Altmetrics approaches also offer speed: indicators of the impact of an article can be measured shortly after the article has been published, which shortens the feedback cycle from years to weeks or even days. The collection of large data using web APIs

makes the process much easier. Altmetrics also offers the transparency on how data is collected from multiple sources. Moreover, altmetrics approaches make it possible to measure online activity associated with non-traditional research outputs and publications. In addition to books and journal articles, altmetrics can also detect activity associated with datasets, software, slides, and blog posts; and the activity that is detected may relate to interest from the general public beyond the scientific community (Piwowar, 2013; Zahedi, Costas, & Wouters, 2014).

According to Neylon (2014), online user interaction data can also be understood as a collection of signals emanating from the unmapped path that information traverses as it travels across knowledge landscapes. Understanding how research is accessed, shared and used by both scholarly communities and the general public has the potential to inform the construction of efficient communication systems, which can maximize the reach and value of investments in research.

Furthermore, altmetrics has the potential to trace research trends. Wang, Wang and Xu (2013) designed a method to detect emerging research trends in real time. In their research, these authors investigated download statistics of articles in the *Journal of Scientometrics*. They aggregated keywords in the downloaded articles and analyzed trends in downloading (Wang et al., 2013).

However, although altmetrics offers many benefits, it also has limitations (Bornmann & Haunschild, 2017; Bornmann, 2014a; Wouters & Costas, 2012). Priem (2014) lists three of these, which are ease of gaming, possible biases, and a lack of theory. Regarding gaming, NISO, in their whitepaper on altmetrics published in 2014, points out that "one important aspect of data quality is the potential for gaming metrics, e.g., behavior that is meant to unfairly manipulate those metrics, generally for one's benefit. Many alternative assessment metrics are more prone to gaming compared to traditional citations" (p.9). It is conceivable, for example, that a researcher could write a script capable of tweeting thousands of tweets about their research. However, as Priem states, traditional metrics are not free from exploitation either.

Nonetheless, the gaming of metrics systems is also being countered by the development of new tools. For example, beginning in 2009, editors at a set of Brazilian journals attempted to artificially increase their journals' impact factors by

citing one another's journals. By 2011, this had increased each journal's impact factor and made each of the journals appear more influential. However, by 2013, Thomson Reuters had implemented a new algorithm to detect more elaborate ways of raising impact factors through "self-citation". The algorithm identified the four Brazilian journals (Van Noorden, 2013).

In an effort to counter the gaming of altmetrics, PLOS also developed a tool called DataTrust, which applies methods such as excluding artificial traffic (Lin, 2012).

Thelwall regarding the ease of manipulation of altmetrics:

> "Altmetrics [also] have the potential to be used for impact indicators for individual researchers based upon their web presences, although this information should not be used as a primary source of impact information since the extent to which academics possess or exploit social web profiles is variable" and that, "more widely, however, altmetrics should not be used to help evaluate academics for anything important because of the ease with which they can be manipulated". (Thelwall, 2014)

Concerning bias issues, Priem finds it acceptable that scientists use the available technology to capture the scientific community's attention by starting a conversation among their peers. He also points out that although altmetrics, as a new area of metrics, currently lacks a body of theory (Taylor, 2013), bibliometrics struggled with the same challenges in its early stages. Priem quotes Garfield:

> We still know very little about how sociological factors affect citation rates... On the other hand, we know that citation rates say something about the contribution made by an individual's work, at least in terms of the utility and interest the rest of the scientific community finds in it (Garfield, 1979).

Taylor states that, as with any system that relies upon measurement-by-proxy, conclusions about what those measurements may mean are only reliable when supported by appropriate theory and evidence. According to him, it took approximately 20 years for bibliographic citation analysis to achieve acceptability as a measure of academic impact (Vaughan & Shaw, 2003), and it may well take another 20 years for web analytics to provide an adequate picture of how scholarly research influences society as a whole (Mike Taylor, 2013). It is important to note

that although bibliometrics and peer review are the accepted approaches for measuring scholarly impact within academia, there is still no recognized framework for measuring social impact (Bornmann, 2014a).

2.4.5 Altmetrics Research

To measure societal impact, altmetrics has moved from case studies at its beginning (Bornmann, 2014a) to population analysis today, using population derived from publication indices including Web of Science, Scopus and Microsoft Academics Graph. Altmetrics offers large volumes of useful data for societal impact measurement; however, it is not yet clear what the individual metrics are measuring. Most of the studies that have empirically investigated altmetrics have focused on correlations between citations and altmetrics (Bornmann, 2015; Costas et al., 2015a; Torres-Salinas et al., 2013).

Such studies have examined correlations between citation metrics on the one hand and new metrics, such as bookmarks in the Mendeley reference manager, mentions on Wikipedia, bookmarks on the social bookmarking web service Delicious, and tweets and recommendations of important articles on the post-publication peer review service site F1000 on the other (Priem, 2014). Researchers have reported a strong correlation between inclusion in Mendeley and WoS (Haustein et al., 2014a; Priem et al., 2012). Nielsen (2007) identified a correlation between citations on Wikipedia and the Journal Citation Report, which is a web-based tool from Thomson Reuters that makes it possible to evaluate and compare journals using citation data. Zahedi and colleagues (2014) investigated the distributions of altmetrics across different academic fields and document types and over publication years. They found a moderate Spearman correlation (r=0.49) between Mendeley readership countrs and citation counts. Waltman and Costas (2013) demonstrated a clear correlation between F1000 recommendations and citations.

In addition to these correlation analyses, new approaches will start to emerge in altmetrics studies. Haustein et al. argue that, as was the case in citation metrics research, there is also a need to define the meaning of the various indicators in altmetrics (Haustein, Bowman, et al., 2015).

2.4.5.1 Correlation analysis in altmetrics

Yan and Gerstein (2011) produced a Spearman correlation matrix of 18 different metrics, including article usage statistics (HTML views, PDF downloads and XML downloads), citation statistics (PubMed, CrossRef and Scopus), blog coverage (Bloglines, Nature Blogs and Postgenomics), social bookmarking (CiteULike) and various online ratings employed on the PLOS website. Their findings show that the number of citations is most strongly correlated with access statistics (with an average Spearman correlation r=0.44 and the highest correlation with the number of PDF downloads r=0.48). This is followed by the number of bookmarks (average Spearman Correlation r=0.2). Among article access statistics, the number of PDF downloads correlates strongly with the number of HTML views (r=0.91, p=0), and these article access statistics are generally aligned with social bookmarking metrics and blog coverage metrics. This suggests that the media coverage of a specific article may contribute to the article's access statistics, or vice versa.

In addition to these investigations, Yan and Gerstein examined the propagation of information using a time series analysis of web accesses. They found that from the first month to the second month following publication, the average number of views declines rapidly, while the rate of decline decreases as time progresses. They argue that the first month to the second month period the number of views was driven by the fame of the article, while the next long-term period was following the same pattern with citation statistics. These authors' findings align with their expected result, which was that older papers receive less attention.

Priem et al. (2012) conducted a correlation analysis on 24,331 PLOS articles to test for a relationship between the number of citations with altmetrics activity. They found from the correlation analysis and an accompanying factor analysis that citation and altmetrics indicators track related but distinct impacts, with neither alone being able to provide a complete overview of scholarly use.

Haustein, Costas and Larivière (2015), in their large-scale study, analysed patterns of five social media metrics with citation data for 1.3 million papers published in 2012 and covered in Web of Science. They found the presence of those papers on social media was low, with 21.5% of papers receiving one or more tweets, 4.7% being shared on Facebook, 1.9% mentioned on blogs, and 0.7% discussed in

mainstream media. Although the presence of papers on social media was low, Costas and colleagues suggested that altmetrics could be complementary to citations in order to inform other types of impact, such as societal or cultural impact, especially in the HSS, which has a greater presence on social media (Costas et al., 2015a). The following sections will discuss altmetrics studies conducted on different social media platforms.

2.4.5.1.1 Wikipedia

Priem and colleagues, in their 2012 study, found that 5% of PLOS articles were cited on Wikipedia. Although correlation strength varied across journals, they found positive correlations (0.1–0.4) between normalized Wikipedia citations and traditional citations. Although a 2005 study published in *Nature* (Giles, 2005) indicated that Wikipedia contents contained some errors, Nielsen (2007) showed that there was a tendency to cite articles in high-impact journals such as *Nature* and *Science*, which shows Wikipedia's increasing ability to serve as a good information organizer for science in general.

Kousha and Thelwall (2017) tested whether citations from Wikipedia to scholarly research outputs had a relationship with citations within the scholarly literature as captured by Scopus. Their study included 302,328 articles and 18,375 monographs in English indexed by Scopus in the period 2005 and 2012. They found that citations from Wikipedia to articles are too rare to draw a conclusion on the relationship between citations from Wikipedia and those from scholarly articles, with only 5% of articles being cited by Wikipedia across all fields. However, they found that one third of monographs have at least one citation from Wikipedia, suggesting that Wikipedia citations can provide extra impact evidence for academic monographs.

2.4.5.1.2 Twitter

Priem et al. (2012) analyzed the correlation between citations to PLOS articles' citations and tweets mentioning these articles. They found that a high visibility of PLOS articles on Twitter attracts many non-scholarly readers as well (Priem et al., 2012). Haustein, Costas and Larivière (2015) in their study found that 21.5% of papers received at least one tweet. A study of 1.4 million biomedical papers, which examined how often the papers were mentioned on Twitter or saved to Mendeley, found a weak correlation between tweets and citations (Haustein, Peters, Sugimoto,

Thelwall, & Larivière, 2014). This study concluded that number of Mendeley readers and number of tweets mentioning an article are two distinct social media metrics that differ from citations.

A study conducted by Snijder (2016) on OA monographs, which included 400 OA monographs published before 2009 by Amsterdam University Press, with 178 in English, and 212 in Dutch, showed that Twitter mentions and citation behavior were weakly related. Snijder found that although OA has a significant influence on book citations, this effect does not necessarily extend to Twitter mentions. However, he concluded that making books freely available had some positive impact on the number of tweets mentioning the books.

2.4.5.1.3 Bookmarking platforms

Social bookmarking websites such as Delicious and CiteULike allow users to store, organize, and search bookmarks of web pages. Users of these services can annotate their bookmarks using informal tags and other metadata, such as titles and descriptions. Priem and colleagues (2012) found that approximately 10% of PLOS articles were bookmarked in Delicious, and the rates of bookmarking of articles over time decay more quickly after publication than website views and PDF downloads.

2.4.5.1.4 PDF Downloads

Shuai, Pepe, and Bollen analyzed 4,606 preprint articles from arXiv.org published between October 2010 and May 2011. They found that Twitter mentions correlated with downloads and early citations (7 months later after submission of the latest paper) (Shuai, Pepe, & Bollen, 2012). However, Snijder (2010), in his study on OA monographs, indicated that there was no distinctive relationship between scholarly books' downloads and citations. However, Snijder noted that the nine-month period of analysis is relatively short for scientific disciplines in which books rather than articles are the norm. In addition, he found that although online usage is higher for fully accessible monographs, this was not reflected in their sales.

2.4.5.1.5 Scholarly Blogging

Scientific communication increasingly occurs on web technology platforms such as scholarly blogs, where users refer to and comment on traditional papers. Scholarly blogs provide a valuable source for measuring instant commentary on publications. Groth and Gurney (2010) applied a webometrics and bibliometrics approach to

chemistry blogs and showed that, compared to academic literature, scientific conversation on the web is more immediate and contextually relevant and has a larger non-technical focus. Shema, Bar-Ilan, and Thelwall stated that although scholarly blog posts are associated with increased visibility and impact, there are some issues with using them as an altmetrics source. For example, blog posts cover only a small percentage of articles, criteria based on which to classify a blog as a scholarly blog are lacking, and there is uncertainty about the sustainability of blogs, which is not the case for journals (Shema et al., 2014).

2.4.5.1.6 Reference managers

Social reference managers such as Mendeley and CiteULike allow users to save, manage, and share scientific literature online. Mohammadi and Thelwall (2014) compared articles' Mendeley readership counts with their citations. They used Mendeley data to discover patterns of information flow between scientific fields and found that Mendeley readership data can be used to provide evidence of a publication's impact at an earlier stage than is possible with citation counts. They also argue that Mendeley has the advantage of covering more types of users, such as undergraduate and postgraduate students as well as practitioners, whereas citation data come only from authors.

Mohammadi, Thelwall, Haustein, and Larivière (2015) investigated the types of readers reading research articles on Mendeley. They used bookmarking counts as an indicator of readership and examined the context in which users were reading the bookmarked articles. They found that after PhD students, postgraduate students and postdoctoral researchers were the two most common types of readers of articles on Mendeley. Moreover, they found a small group composed of readers outside of higher education institutions. According to these authors, reference manager data can reveal aspects of the readership of research articles. This is especially important in the disciplines for which citation-based indicators are least reliable, such as the humanities and social sciences (Mohammadi, Thelwall, Haustein, et al., 2015; Mohammadi, Thelwall, & Kousha, 2015).

2.4.6 Interpreting altmetrics indicators using social theories

As with citation theories, which aim to understand authors' motivations for citing a document, altmetrics researchers started to investigate users' motivations for mentioning a research output on social media. Following Garfield's (1965) article "Can citation indexing be automated?", Bornmann listed 15 possible motivations for citing, which are as follows:

1. Paying homage to pioneers;
2. Giving credit for related work (homage to peers);
3. Identifying methodology, equipment, and so forth.;
4. Providing background reading;
5. Correcting one's own work;
6. Correcting the work of others;
7. Criticizing previous work;
8. Substantiating claims;
9. Alerting readers to forthcoming work;
10. Providing leads to poorly disseminated, poorly indexed, or uncited work;
11. Authenticating data;
12. Identifying the original publications in which an idea or concept was discussed;
13. Identifying the original publication or other work describing an eponymic concept or term (…);
14. Disclaiming the work or ideas of others (negative claims);
15. Disputing the priority claims of others (negative homage).

Following Taylor's findings (2013), Bornmann added to this list six more motivations, which are especially relevant for mentions of papers on social media. These motivations are:

1. Building a network of related researchers;
2. Building a reputation as a good networker;
3. Paying visible homage to a senior researcher;
4. Seeking the attention of a senior researcher;
5. Demonstrating that one's reading is up to date;
6. Intimidating critics with the breadth of one's reading.

Bornmann states that not only should robust and reliable methods for societal impact measurement be developed, but the connection between these measurements and sociological theories, theories of informetrics, and theories of scholarly communication should also be extensively studied (Bornmann, 2016). In an effort to understand the behaviors observed on social media, Haustein and colleagues (2015) launched a discussion on approaches for interpreting altmetrics, drawing on both citation and social theories. Although correlation analysis may make it possible to predict the impact of a particular research output more quickly, Haustein and others have gone further using different types of data, by attempting to explain behaviors observed in relation to scholarly outputs on social media. Their approach seems likely to be helpful in the development of an altmetrics theory. Haustein et al. classified the types of acts, or behaviors, observed in relation to scholarly publications on social media into three categories:

- 'accessing' for viewing and downloading
- 'appraising' for mentioning articles on various platforms, such as blogs, Wikipedia and social networks
- 'application' for adapting and transforming theories, frameworks, methods, or results from an article, a scientific document, software code, etc.

In order to interpret these acts, Haustein et al. first used citation theories such as the normative theory, the social constructivist theory, and the concept symbol theory, which were discussed at the beginning of this chapter. In applying these theories to altmetrics studies, they found that the Mertonian norms were identifiable in relation to reviewing and recommending published works via the F1000 site. They found that the Mertonian norms were less applicable to citations on blogs, and that these norms did not satisfactorily explain patterns of Twitter mentions. Using the social constructivist theory, they found that on Mendeley and Twitter, documents with more interactions on social media gained higher visibility, which was taken as evidence of the Matthew effect. They suggested that Twitter data might be understood as providing insight into the public reception of research and can therefore serve as a mechanism for capturing the public perceptions of scientific concepts, rather than the scientific community's response to a publication.

In addition to citation theories, Haustein et al. proposed three social theories for interpreting social media actions relating to research outputs, including social capital, attention economics, and impression management. In using these theories, they recognized the inherently social nature of the platforms from which altmetrics data are drawn. In the following subsections, these three social theories are reviewed.

2.4.6.1 Social Capital

According to social capital theory, actors in networks establish and maintain relationships with other actors in the hope that they will benefit in some way from these relationships. Pierre Bourdieu, in his chapter "The forms of capital", distinguished three types of capital, which are economic, cultural and social (Bourdieu, 1986). Bourdieu defines social capital as "the aggregate of the actual or potential resources which are linked to possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition". For Huysman and Wulf (2004), social capital is "the network ties of goodwill, mutual support, shared language, shared norms, social trust, and a sense of mutual obligation that people can derive value from. It is understood as the glue that holds together social aggregates such as networks of personal relationships, communities, regions, or even whole nations". Ellison, Steinfield and Lampe (2007) state that the resources derived from these relationships can differ in form and function based on the relationships themselves. Social capital enables individuals to draw on resources from other members of the networks to which they belong.

Grannovetter defines the strength of a tie as a combination of the amount of time, emotional intensity, intimacy (mutual confiding), and reciprocal services that characterize the tie (Granovetter, 1973). A "weak tie" is a loose connection between individuals who may provide each other with useful information or new perspectives, but typically not emotional support.

Researchers have emphasized the importance of Internet-based linkages for the formation of weak ties, which serve as the foundation of bridging social capital. Bridging social capital is a type of social capital that describes connections that link people across social groups. It is now possible that new forms of social capital and relationship building will occur on online social network sites. Bridging social

capital may be augmented by such sites, which support weak social ties by allowing users to create and maintain larger and more diffuse networks of relationships from which they could potentially draw resources (Ellison, Steinfield, & Lampe, 2007).

Ellison and colleagues' (2007) study on the benefits of Facebook friends showed a significant connection between Facebook usage and indicators of social capital. They stated that such connections could have notable benefits in the form of jobs, internships, and other opportunities. Haustein et al. (2015) used social capital to interpret users' behaviors on social media. They stated that when a scholar tweets about a scientific document, they are making weak connections with their readers. These authors suggested that when the same user tweets about publications from the same author(s), the connections between the two (or more) authors can strengthen and yield later benefits in the form of collaboration or a letter of reference.

2.4.6.2 Attention economics

Davenport and Beck (2001) define attention as "*...a focused mental engagement on a particular item of information. Items come into our awareness, we attend to a particular item, and then we decide whether to act*".

Attention economics understands human attention as a scarce commodity in an increasingly abundant information environment. Since the limiting factor in the consumption of information is attention, attention becomes increasingly valuable. Herbert A. Simon, who coined the concept of attention economics, stated that what information consumes is in fact the recipient's attention. Thus, there is a need to allocate attention efficiently across the abundance of information sources that might consume it (Simon, 1971).

Researchers have used attention economics to understand user behaviours on social media. Rui and Whinston, who analysed more than 3 million Twitter users, stated that, contrary to Facebook, which is based on real-world friendship, Twitter is effective at connecting people's needs for information and attention. This feature made Twitter better suited to Rui and Whinston's study on attention economics. In their study, they found that users produced information to attract attention, and they contributed attention while consuming information (Rui & Whinston, 2012).

2.4.6.3 Impression management

The process of impression management takes place on highly networked social media, when a user's actions may be motivated by the desire to present a particular image of themselves and involve a desire to avoid shame and embarrassment. Goffman (1978) states that we present ourselves by using impression management, where we reveal certain aspects of ourselves while hiding others. According to him, these behaviours consist of expressions given, such as spoken communication, and expressions given off, such as nonverbal communication cues. Thus, we act strategically to convey an impression to others that is in our interest. We also tend to engage in self-enhancement, in which we attempt to portray ourselves in an overly positive light, even when we interact with strangers (N. Ellison, Heino, & Gibbs, 2006). In their study of Facebook, Gosling, Gaddis, and Vazire (2007) organized group meetings where participants rated themselves and their friends. Eight months later, they compared these ratings to the participants' ideal-self ratings, as well as ratings of how they believed they were viewed on the basis of their Facebook profile. They found that participants enhanced their own self-presentation regarding their personality traits, such as their emotional stability and their openness to experience. In addition to profiles users also enhance their self-presentation in the contents they post. Regarding impression management on Twitter, Gilpin (2011) wrote "Interaction thus plays an especially strong role in identity construction in a conversational medium such as Twitter followers will primarily draw conclusions based on the content of tweets, as well as indications of the intended recipients of those messages".

Since there are different kinds of act, Haustein et al. encourage the thoughtful interrogation of motivations behind user acts on social media observed in altmetrics studies, using methods such as interviews (Mohammadi, Thelwall, Haustein, et al., 2015) and content analysis (Shema, Bar-Ilan, & Thelwall, 2015). This research draws on the approach developed by Haustein et al. (2015), in conjunction with other relevant theories and models used in social media research (Ngai, Tao, et al., 2015), to interpret the collected data on OA monographs.

2.5 Monograph usage

In contrast to journal articles, little work has been done that applies altmetrics methodologies to monographs. Monographs are commonly produced in the HSS rather than the hard sciences. Since HSS disciplines have placed much less emphasis on citation analysis and impact factors than has been the case in the hard sciences, the demand for an alternative to citation analysis has not been as strong among these communities, until recently (Montgomery, 2013).

The high level of demand for new insight into the ways in which scholarly publications are being used, shared and discussed online led Springer Nature to partner with Altmetric.com to develop a platform that would provide book-level and chapter-level metrics. This platform called Bookmetrix was designed to offer altmetrics services for Springer Nature's ebook collections only. Since none of the titles in this study were published by Springer Nature, the Bookmetrix platform was not used.

In 2017, to coordinate and pool university-led scholarly communication activities in Europe, particularly in the HSS, the Open Access in the European Research Area Through Scholarly Communication (OPERAS) programme was started. Funded by the European Union's Horizon 2020 research and innovation program, the OPERAS project aims to develop open scholarly communication by establishing common good practice standards for digital OA publishing, infrastructures, and services (OPERAS, 2018).

To address the long-term requirements of the OPERAS network, the OPERAS-Design (OPERAS-D) project was launched in 2017. The main objective of OPERAS-D is to prepare a design study that defines governance models and scientific and technical concepts for future services for open access publications in the HSS. As part of this project, KU Research, in cooperation with OPERAS partners, published a report on the visibility of OA monographs in the European context. This report explored the extent to which OA monographs are visible to the communities that might make use of them. As part of this study, KU Research investigated OA monographs from six repositories/publishers from six different countries. They discovered that the visibility of books in catalogues varied across publishers and that the variable quality of book metadata created challenges in data

aggregation and analysis (Neylon, Montgomery, Ozaygen, Pinter, & Saunders, 2018).

With the aim of integrating OA monographs into the open science ecosystem in a systematic and coordinated way, the High Integration of Research Monographs in European Open Science (HIRMEOS, http://www/hirmeos.eu) project was developed. HIRMEOS, which is the proof of concept project of the OPERAS project, involves five publishing platforms. One of their objectives is to develop metrics services that will provide altmetrics and citation metrics, as well as a widget with which to display these metrics on partners' websites (HIRMEOS, 2017).

Snijder (2016), one of the few studies conducted on the usage and citations of OA monographs, found that a monograph's OA status had a positive influence on its citations. Making monographs freely available had a clear positive effect on usage: free books were used more than a control group of books that were not OA. This higher usage translated into a higher uptake on social media. Snijder states that Twitter mentions and citation behaviour were only weakly related. Moreover, the likely reasons for OA significantly influencing book citations do not necessarily apply to Twitter mentions, as is the case for journal articles (Adie, 2014; Wang, Liu, Mao, & Fang, 2015). Nonetheless, it is possible to say that making books freely available has some positive impact on the number of tweets relating to these books (Snijder, 2016).

The Crossick Report (2015) mentions the impression that present-day scholars lack the time to read books thoroughly, and it is feared that the academic skill of 'deep reading' may become, or have already become, devalued or lost. In 2014, a survey conducted by OAPEN-UK with UK researchers in the HSS revealed that most of the survey respondents had read their last book for research, writing, teaching or presentation purposes, and that they were very unlikely to have read the whole book (OAPEN-UK, 2014). Instead, they tended to read specific chapters. However, respondents who had read for the purpose of writing a book review were much more likely to have read the whole book. Annotation services such as Hypothes.is or PaperHive can provide useful data with which to obtain new insights into the ways in which research communities are reading scholarly monographs. Therefore, combining usage data and social media data will allow us to construct a uniquely

detailed picture of how and by whom OA books are being accessed and used (Adie & Roe, 2013; Padula & Williams, 2015; Williams, 2015).

To track and analyze the impact of monographs, Nederhof suggests that a "citation window" of a period of six to eight years is most suitable, because this period will allow for the monograph's worldwide reception to be captured and reflected in the results (Nederhof, 2011; Snijder, 2016). Since the KU titles are relatively new, the present study will also shed light on whether the relatively short period from the books' date of publication presents a challenge for correlation and altmetrics analyses of OA monographs.

2.6 Conclusion

Open access monographs and journal articles differ significantly in the ways in which they are distributed, shared and mentioned. First, monographs can be hosted in multiple repositories and be distributed in different ways, such as in whole-book or single-chapter format. These scholarly books can be in many formats as well, including PDF, ePub, MOBI and HTML. They are also identified in different ways, using ISBN and DOI identifiers, and one book is likely to have multiple ISBNs and DOIs. Unlike journal articles, monographs' citation coverage is poor. These differences make it considerably more difficult to track scholarly books on social media.

To date, with the exception of the webometrics approach, most of the existing studies on citation and altmetrics have focused on journal articles. However, a great deal of work in this domain needs to be conducted on scholarly books as well, since books and journal articles are different types of research outputs. Moreover, the focus of altmetric studies has mainly been on correlation analyses between various social media data sources and citation data. There is a significant gap in the existing literature that tracks data related to OA monographs.

The purpose of the present study is to understand the complete journey of a scholarly OA monograph. Each aspect of a book's digital life, such as its discoverability, web visibility, usage, social media mentions, and citations in conventional academic publications will be studied together. Bearing the unique characteristics of and challenges related to scholarly books in mind, this study

combines different approaches covered in the literature to track monographs' journeys. In so doing, the online dissemination of OA monographs can be understood in a holistic way.

There are aspects about the monograph that need to be explored using a combination of approaches. The main gap that needs to be addressed is the interpretation of social media data. Literature on this subject is not only absent in the monograph realm, but in the scholarly context more broadly. Haustein and colleagues (2015) have begun the work of interpreting social media data relating to journal articles using citation and social theories. However, much work remains to be done in this area. The interpretation of acts on social media related to research outputs can provide new insights into how to make platforms intended to enhance scholarly communication more useful and transparent.

This study therefore addresses important gaps in our existing knowledge about the different phases of OA monographs' journeys across the digital landscape, as well as about the relationships between the interactions that comprise these journeys and their interpretation. The study also contributes by aiming to unify the interpretations of these acts, which are not only limited to social media behaviors, but include all actions related to scholarly research outputs, including citations.

The following chapter presents a comprehensive review of the literature on the discoverability, visibility and usage of OA monographs. It first tracks and analyses the data around these three aspects and subsequently interprets these data.

3 Discoverability, visibility and access of open access monographs

3.1 Introduction

This chapter examines the 28 KU titles' discoverability, visibility and access in the digital landscape. A range of webometrics approaches were used to identify the visibility of these titles on websites, blogs, and social networks. Access statistics were employed to investigate the locations and institutional backgrounds of the readers of these titles and to determine how they had discovered these books.

The chapter is divided into three sections: discoverability, visibility, and access. The discoverability section explains how OA monographs are hosted and recorded in repositories and how these records are shared on the Internet so that search engines and library catalogues can index them. The second section investigates the visibility of the 28 KU titles on the web. Webometrics methods are used to identify the numbers and types of sites that mention the titles and the ways in which they mention them. The third section considers different methods for collecting repository platforms' access statistics, including downloading access reports, scraping from platform pages, and using web analytics services such as Google Analytics. This section discusses the usefulness of these access statistics in identifying readers' locations and the ways in which monographs are discovered. It also identifies effective channels for disseminating information about these monographs. Different kinds of access statistics are used in order to understand the causes of high or low monograph access and to draw conclusions regarding institutional access to these monographs. The findings on visibility and access are presented and discussed in their respective sections. At the end of this chapter, a case study title is selected from the 28 titles, and the approaches reviewed in the visibility and access sections are applied to this title.

This chapter uses low-cost, widely available tools and services for collecting access data. These tools and services are relatively easy for publishers or repositories to use

to obtain an overview of their titles' performance. They can also contribute to identifying effective channels to promote books, such as email or social media campaigns, and to measuring the results of these efforts.

3.2 Datasets

Four datasets were used to examine the visibility and access of the 28 KU pilot collection titles (Table 3.1). These datasets include three types of data: visibility data, repository access data, and web access statistics.

Table 3.1: Datasets used in this chapter.

Type	Detail	Date
Visibility	URLs of all web pages on which the 28 titles and author names were present	July 2017
Visibility	Links on the web pages on which the 28 titles and author names were present	July 2017
Access	OAPEN repository access data	March 2014 – June 2017
Access	OAPEN web access statistics	March 2014 – June 2017
Access	HathiTrust web access statistics	March 2014 – June 2017

3.2.1 Visibility data

Using the list of 28 book titles and author names shown in Table 3.2, the URLs of all the web pages on which these title and author names were present were collected for webometrics analysis in July 2017. The host names and domain names were extracted from the URLs and counted for each title.

Table 3.2: List of 28 titles from the KU pilot collection.

Title	Subtitle	Creator
1. Fighting for a Living		Zürcher, Erik-Jan
2. Law, Liberty, and the Pursuit Of Terrorism		Douglas, Roger
3. Thinking and Killing	Philosophical Discourse in the Shadow of the Third Reich	Segev, Alon
4. China's iGeneration	Cinema and Moving Image Culture for the Twenty-First Century	Johnson D., Matthew; Wagner B., Keith; Yu, Tianqui; Vulpiani, Luke
5. The Myth of Piers Plowman	Constructing a Medieval Literary Archive	Warner, Lawrence
6. Governing Failure	Provisional Expertise and the Transformation of Global Development Finance	Best, Jacqueline
7. The Emergence of Irish Gothic Fiction	Histories, Origins, Theories	Killeen, Jarlath
8. Partisan Gerrymandering and the Construction of American Democracy		Engstrom J., Erik
9. Constructing Muslims in France	Discourse, Public Identity, and the Politics of Citizenship	Fredette, Jennifer
10. Ever Faithful	Race, Loyalty and the Ends of Empire in Spanish Cuba	Sartorious, David
11. Rhetorics of Belonging	Nation, Narration and Israel/Palestine	Bernard, Anna
12. Biological Relatives	IVF, Stem Cells and the Future of Kinship	Franklin, Sarah
13. Electronic Iran	The Cultural Politics of an Online Evolution	Akhavan, Niki
14. Verse and Transmutation	A Corpus of Middle English Alchemical Poetry	Timmermann, Anke
15. Understanding the Global Energy Crisis		Coyle D., Eugene; Simmons A., Richard
16. In Search of the Amazon	Brazil, the United States and the Nature of a Region	Garfield, Seth
17. Passionate Amateurs	Theatre, Communism and Love	Ridout, Nicholas
18. The Ethics of Armed Conflict	A Cosmopolitan Just War Theory	Lango W., John
19. Making and Unmaking in Early Modern English Drama	Spectators, Aesthetics and Incompletion	Porter, Chloe
20. Composing the Party Line	Music and Politics in Early Cold War Poland and East Germany	Tompkins G., David
21. Aging Gracefully in the Renaissance	Stories of Later Life from Petrarch to Montaigne	Skenazi, Cynthia
22. My Voice Is My Weapon	Music, Nationalism and the Poetics of Palestinian Resistance	McDonald A., David

Title	Subtitle	Creator
23. Human Rights and Democracy	The Precarious Triumph of Ideals	Landman, Todd
24. Beastly Journeys	Travel and Transformation at the fin de siècle	Youngs, Tim
25. Networks and Institutions in Europe's Emerging Markets		Schoenman, Roger
26. Oaths and Swearing in Ancient Greece		Sommerstein H., Alan; Torrance C., Isabelle
27. On Global Citizenship		Tully, James
28. The World Jewish Congress During The Holocaust	Between Activism and Restraint	Segev, Zohar

3.2.2 OAPEN repository access data

COUNTER-compliant access reports were collected from the IRUS-UK website for the period of March 2014 to June 2017. Two types of access reports were downloaded for the KU pilot collection titles: country-based access reports and IP address-based access reports. Country-based access reports cover country-specific monthly downloads for each title. Instead of countries, IP addresses are used in the IP address-based access reports. Location and institution names were included for each IP download record in the IP access reports for the KU pilot collection. The OAPEN metafile was used to identify other titles in the repository for comparison. Then, for these identified titles, the country-based access reports for specific time periods were downloaded.

3.2.3 Web access statistics

The monthly access statistics of the book presentation web pages of the 28 KU titles in the OAPEN repository were collected for the period of March 2014 to June 2017. This access data includes unique page views, referring web pages, and social network sources for the 28 titles' presentation pages. HathiTrust web access statistics were also collected for the period of March 2014 to June 2017, since these provide access statistics for the HTML versions of the titles. In conclusion, the datasets made it possible to analyse the geographic information related to each title. The access datasets also included monthly granularity over the period of March 2014 to June 2017. This made it possible for the access investigation to include a time scale component.

3.3 Discoverability of monographs

Before examining the discoverability of titles, it is worth reviewing how repositories host monographs. Open access monographs are hosted in different repositories, which have different purposes and work in different ways. These differences affect how titles are discovered. The following section discusses how monograph records in repositories are disseminated on the web.

3.3.1 Repositories

To store and disseminate scholarly information such as digital collections of books, papers, theses, media, and other works, organisations use digital repositories. These repositories are of various types, such as subject-based repositories, research repositories, national repository systems, and institutional repositories (Armbruster & Romary, 2012). These types depend on the users, hosting institution and the repository's purpose. However, using Armbuster and Romary's classification, it is difficult to classify repositories into just four types, because the boundaries between repository types are unclear. Sometimes, it is difficult to distinguish a repository from a digital library or even a publisher platform. In addition, some of these platforms offer similar services. In this study, platforms that host monographs are referred to as "repositories". However, the purpose of a repository can determine the discoverability of the titles it hosts. This study's focus is on OA repositories that host monographs in the HSS, such as OAPEN and HathiTrust.

In fact, there are many OA repositories. According to the Registry of Open Access Repositories (ROAR), which promotes the development of OA by providing up-to-date information about the growth and status of repositories around the world, there are more than 4,650 open repositories globally (ROAR, n.d.). According to OpenDOAR, another authoritative directory focussed on academic OA repositories that is developed and maintained by the University of Nottingham, as of November 2018, at least 3,780 of these repositories were academic OA research repositories (Millington, 2006). The following section reviews how the KU pilot collection titles are discovered on these OA repositories.

3.3.2 Discoverability of Titles

There are two factors that affect the discoverability of monographs. One is indexing titles by search engines, and the other is registering titles to book directories. In general, library catalogues implement these search engines or directories in their catalogues.

3.3.2.1 Search engines

Although the details of how most search engines index titles are not clear, in general, they usually download a file from a repository that lists all the hosted titles. These titles can also be gathered automatically from the repository using an internationally agreed-upon set of technical standards with which OA repositories are intended to comply. This means that repositories provide the metadata (bibliographic details such as author name, institutional affiliation, date, article title, abstract, etc.) of each item they contain in the same basic format. In this way, they are interoperable and cross-searchable by other repositories. The common protocol to which open access repositories adhere is called the "Open Archives Initiative Protocol for Metadata Harvesting" (OAI-PMH).

The OAI-PMH specifies how metadata are structured and presented for harvesting by external services. Metadata following this protocol are encoded in extensible markup language (XML) format. An XML-encoded OAI-PMH record is organised into the following parts:

- Header: Unique identifier, datestamp, set membership, status (optional).

- Metadata: Set of metadata, often in simple Dublin Core which is a set of terms that can be used to describe resources.

- About: Optional rights statements, provenance, and other uses.

A specific type of OAI-PMH, which is used for library bibliographic data, is the OAI-Dublin Core (DC) metadata schema. This is the type that many library vendors employ for exposure and harvesting (OpenAIRE, n.d.).

Using OAI-PMH protocol, repository contents can be indexed by search engines, which assist in the creation of online OA databases of research from across the globe. There are academic search engines that provide specific access to scholarly

Internet resources, including monographs. One important example is the Bielefeld Academic Search Engine (BASE), an OA academic search engine created by the Bielefeld University Library in Germany. BASE also offers services to companies like EBSCO, which provides the EBSCO Discovery Service (EDS) to libraries (Price, 2015). BASE is founded on the free and open-source library search engine software VuFind. Like Google, BASE harvests OAI metadata from institutional repositories and other academic digital libraries that implement OAI-PMH and then normalises and indexes these data for searching.

Another search engine that uses OAI-PMH is OpenAIRE. OpenAIRE was created by the European Commission (EC) to support the implementation of EC and European Research Council (ERC) OA policies. OpenAIRE is a network of repositories, archives, and journals that support OA policies. As of November 2017, OpenAIRE had collected approximately 23 million documents from 980 compatible data providers. OpenAIRE aims to promote open scholarship and substantially improve the discoverability and reusability of research publications and data (OpenAIRE, n.d.).

3.3.2.2 Directory of Open Access Books

The other discovery service that library catalogues use for OA monographs is the Directory of Open Access Books (DOAB). The DOAB is maintained by the OAPEN Foundation and is based at the National Library of the Netherlands. As of July 2017, the DOAB provides a searchable index and links to the full texts of 7,814 academic peer-reviewed books and chapters from 205 publishers. The DOAB covers multiple subject areas and has specific requirements for the inclusion of books in its directories. All books listed in the DOAB have an OA licence, and collaborating publishers are screened for their peer review policies.

According to McCollough (2017), the aggregation of OA monographs' metadata by a trusted entity such as the DOAB plays a significant role in facilitating OA book discoverability in library catalogues. According to the former University of California Press director and current CEO of PLOS Alison Muddit (McCollough, 2017), there are two key challenges associated with making OA content fully discoverable within library catalogues; a task that is crucial for increasing access and impact. The first challenge is that some librarians see their cataloguing role as

pertaining only to their local collections, which greatly limits discoverability in a global networked environment. The second challenge is that there is no easy way for OA publishers to produce records through vendors, because vendors' revenue model depends on being able to take a portion of a book's sales price (McCollough, 2017). Therefore, the aggregation of OA metadata by a trusted entity such as the DOAB plays an especially important role in facilitating OA book discoverability in library catalogues. According to McCollough, research titles that are not registered in the DOAB are much less discoverable in library catalogues than their registered counterparts.

All of the 28 KU pilot collection titles were hosted in OAPEN, and 24 titles were registered in the DOAB. Since the BASE and OpenAIRE search engines index both the OAPEN repository and the DOAB, all 28 titles were indexed by these search engines. In the following section, the visibility of these 28 titles is reviewed by examining the presence of the title names on web resources.

3.3.3 Discussion

The main issue regarding the discoverability of the 28 KU titles was poor metadata. My research as a member of the KU Research team, which in turn formed part of the OPERAS-D project, shows that there are many issues with the metadata of repositories which reduce books' discoverability. During the OPERAS-D project, we found that titles may differ across different repositories, ISBN or DOI identifiers may be incorrect or missing, and titles' categories may be too general (Neylon et al., 2018). There were also differences in the metadata schemata, and different repositories used different metadata formats and included different content. It is important to describe items sufficiently in repository metadata, so that search engines such as OpenAIRE, BASE, and Google Books can index their content.

There were also other issues relating to the discovery of monographs. For example, when a new edition or new translation of a book is released, the question arises as to how the content is distributed to repositories. Should the ISBN and DOI be changed? Should a DOI and ISBN be assigned to each format of the book? These issues and challenges are discussed in detail in the final chapter of this study.

3.4 Visibility of monographs

Webometrics methods were used to capture the KU pilot collection titles' visibility on the web and to understand the role of web visibility in the dissemination of and access to these titles. Both qualitative and quantitative techniques were used. Using only quantitative techniques would risk the results being superficial or misleading (Thelwall, 2009). The roles of geography, type of content, and number of links and web pages in the dissemination of these books were investigated. The study identified types of websites that trigger conversations and direct traffic to repositories, such as news sites, academic sites, blogs, and journal sites. Webometrics methods were used for web impact assessment, which in this thesis is referred to as "web visibility assessment", as here it indicates the titles' presence on the web. Three different webometrics methods were used for the web visibility assessment. These are web mentions, content analysis, and URL analysis techniques. To prevent confusion in relation to the use of the term "mention" in the following chapter on social media, web mentions are here referred to as "web presence".

3.4.1 Web presence

Web presence is measured by submitting a collection of terms or document titles to a search engine and then reporting the hit count estimates as an indication of visibility. In this study, instead of using an estimated number, the full results number is used. To search for a book on a search engine, Thelwall suggests using the title in quotation marks and adding the author's surname to the search string as a refinement (Thelwall, 2009). However, in this study, some of the monograph titles were short and generic, such as the title "Governing Failure", which had an author with the surname of "Best". Submitting these strings returned a large number of irrelevant results, such as web pages on 'governing failures' in different fields using the 'best' approaches, amongst others. To return results related to the book, the query had to be more precise. The first strategy for addressing this challenge involved adding the author's first name, as in

```
"Governing Failure" Jacqueline Best
```

This approach was chosen because, if the author's name and surname were inside the quotation marks, pages that used just the initial of the first name or pages that put

the first name after the surname would be missed. If the title were placed outside of the quotation marks, pages on failure in governing that had no relation to the book or its subject would be obtained. In these pages, for example, the strings “Jacqueline” and “best” may have been encountered in the text. Thus, subtitles (“Provisional Expertise and the Transformation of Global Development Finance”) were also added to these search strings. In this case, the search string was transformed to

```
“Governing Failure” “Provisional Expertise and the Transformation of Global Development Finance” Best.
```

Therefore, for short and generic titles, subtitles, if present, can also be used to retrieve more relevant results.

For the webometrics analysis, Webometrics Analyst 2.0 software was used to collect and analyse the data (Thelwall & Sud, 2012). Because of Google’s access restrictions, Webometrics Analyst uses the Microsoft Bing search engine through Microsoft’s Azure API.

3.4.2 Content analysis

In order to categorise the web resources obtained from the web presence analysis, a content analysis was conducted. To include an academic angle, Thelwall adds citation data, which can be obtained from services such as Google Scholar (Thelwall, 2009). Google Scholar citation data is analysed separately in the following chapter.

The objective of the content analysis was to determine the creator and the purpose of the citing page. Thelwall suggests that this analysis should be used to complete sentences like: “Title X was mainly mentioned online by A and B” or “Title X was mainly mentioned online because of A and B”. Since non-academics read HSS books as well, web resources belonging to organisations outside of universities, including booksellers and journals, were also categorised, in addition to blogs and news sites. This categorisation was helpful for investigating the audience reach of titles. Because of time limitations and because the aim was solely to obtain context rather than to accurately distinguish between categories, an informal content analysis was conducted using intuitive judgments of categories (Neuendorf, 2002). Since more than 3,200 web resources citing KU pilot collection titles had to be

categorised, a sampling method was employed to check and classify these resources. In order to obtain an appropriate spread of results and prevent introducing a systematic bias, every 10th resource (i.e. 10th, 20th, 30th, etc.) in the search results was checked and categorised.

3.4.3 URL analysis

After collecting, analysing and categorising web resources in which title names were present, information was extracted from the URLs of the resources returned by the Bing search engine. Extracting the top-level domain (TLD) from a website's domain name makes it possible to identify the country that hosts the web resource. One of the drawbacks of this analysis is the prevalence of generic TLDs such as .com, which are not country-specific. URL analysis is useful in order to identify the sites on which the 28 title names occur most frequently. Unique domain names were then extracted from these resources' URLs. Titles were sorted according to the number of unique domains in which they were present. Lastly, the second-level domains (SLD) of the relevant resources were also analysed to determine whether they belonged to university sites. The TLD refers to the last segment of a domain name, or the part that follows immediately after the period, and the SLD is the domain directly below the top-level domain (TLD). For example, for the host ccat.curtin.edu.au, the TLD is 'au', indicating that the country of the host is Australia, the SLD is 'edu.au', indicating that the host is an Australian educational institution, and the domain is the curtin.edu.au, indicating that the resource belongs to Curtin University in Australia.

3.4.4 Repository presence

As opposed to articles, which are normally hosted on one publisher's site, monographs are often found on multiple repositories. Since there are many repositories, it is important to determine which of these repositories is more visible in disseminating the 28 KU monographs. To determine the online visibility of these repositories for the 28 titles, the URLs of all the resources in which the title names were present were investigated, and the links inside them were collected. Using the links collected, the unique domain names were identified, and the occurrences of the repositories' domain names were examined.

3.4.5 Findings

The findings regarding the visibility of 28 KU monographs obtained through the web presence analysis, content analysis, and URL analysis are presented below.

3.4.5.1 Web presence

To evaluate the visibility of the 28 KU titles on the web, Webometrics Analyst 2.0 software was used. Using this software, the search results delivered for each title by the Microsoft Bing search engine were counted. The title of the monograph in quotation marks was entered as the search string, and the surname of the author was added as follows:

```
"Title_of_the_book" Surname_of_the_author
```

Table 3.3 presents an overview of the Webometrics Analyst results for the 28 titles from the KU pilot collection.

Table 3.3: Overview of Webometrics Analyst results for the 28 titles from the KU pilot collection.

Base query	URLs	Hosts	Domains	SLDs	TLDs
"Biological Relatives" Franklin	253	200	173	26	22
"Governing Failure" Best	248	198	186	30	25
"Ever Faithful" Sartorious	165	122	105	18	17
"Law, Liberty, and The Pursuit Of Terrorism" Douglas	163	120	107	22	20
"Constructing Muslims in France" Fredette	162	115	108	22	21
"My Voice Is My Weapon" McDonald	150	121	103	18	18
"The Ethics of Armed Conflict" Lango	142	102	95	26	24
"The Myth of Piers Plowman" Warner	138	100	93	26	23
"Composing the Party Line" Tompkins	133	97	90	17	16
"The Emergence of Irish Gothic Fiction" Killeen	131	95	84	18	15
"Human Rights and Democracy" Landman	127	99	90	14	13
"Oaths and Swearing in Ancient Greece" Sommerstein Torrance	125	93	88	24	23
"Fighting for a Living" Zürcher	117	93	83	21	19
"Understanding the Global Energy Crisis" Coyle Simmons	116	83	70	19	17
"Beastly Journeys" Youngs	115	86	79	21	18
"Passionate Amateurs" Ridout	112	82	76	16	15
"On Global Citizenship" Tully	110	83	79	17	15

Base query	URLs	Hosts	Domains	SLDs	TLDs
"Partisan Gerrymandering and the Construction of American Democracy" Engstrom	109	84	78	22	21
"Networks and Institutions in Europe's "Emerging Markets" Schoenman	100	75	72	21	19
"Aging Gracefully in the Renaissance" Skenazi	93	69	67	23	22
"Rhetorics of Belonging" Bernard	92	68	62	15	12
"Making and Unmaking in Early Modern English Drama" Porter	86	65	62	16	15
"Thinking and Killing" Segev	85	66	60	18	17
"The World Jewish Congress During The Holocaust" Segev	68	46	45	14	13
"In Search of the Amazon" Garfield	48	41	39	10	10
"Electronic Iran" Akhavan	37	35	35	12	11
"Verse and Transmutation" Timmermann	37	29	28	7	7
"China's iGeneration" Johnson Wagner Yu Vulpiani	7	4	4	2	2

The “URLs” column in Table 3.3 shows the number of web resources in which the monograph’s title and author’s surname are present. The “Hosts” column shows the number of unique host names extracted from the URLs of the web pages on which each title name is present. Because of the unregulated nature of the web, some websites may have mirrors or copies in different hosts under the same domain. However, host information is still valuable to provide a basic indication of these titles’ visibility on the web. Because the same title under different context can be present under the same domain in different hosts, which would increase their visibility. This is why the results were filtered according to unique hosts for each title.

Unique domain names from these hosts were also extracted and counted. These numbers are shown in the “Domains” column. Some domains, such as those belonging to publishers, funding institutions or universities, may have different subdomains (hosts or websites). These subdomains may belong to a faculty, an online journal, a blog, or a press website, each of these hosts in which the titles are present. Therefore, counting each separate domain helps to identify the number of entities that host these subdomains.

3.4.5.1.1 Most and least present titles on the web

Table 3.3 shows that the two title names most present on the web are Franklin's "Biological Relatives", with 253 URLs, and Best's "Governing Failure", with 248 URLs. These title names' web presences are much higher than those of the other title names, which have fewer than 150 results each. Some of the resulting URLs featured content irrelevant to the monographs in question, especially in the case of Best's "Governing Failure", because this search string is short and composed of common words. For this reason, both titles and subtitles were included in the query.

One of the challenges involved in using subtitles in the search string was that web resources in which only the book's title was present were overlooked in the results. It was observed that most of the web resources only included the title name and the author's surname. On the other hand, if only titles were used in the search string, it is likely that many unrelated web resources would be found for titles composed of common words. Thus, for many titles, the subtitle does not have to be included in the search string. However, when relatively few titles are being investigated, as in the case of this study, it is advisable to check for titles with an abnormal number of presences. Therefore, "Governing Failure" and "Biological Relatives" titles were checked by adding their subtitles to the search string. The number of results for the "Governing Failure" title fell from 248 to 38 URLs, whereas the "Biological Relatives" title only fell from 253 to 222 pages, as shown in Table 3.4. For the "Governing Failure" title, 21 of the 210 discarded URLs were examined, and it was found that only three URLs were related to the book without using the subtitle. The other URLs were mainly related to engineering, where the term "governing failure modes" was used, and to governing failures in other subjects. Two pages on booksellers' sites were related to other books, and there was no trace on these sites of Best's "Governing Failure". This was probably due to the fact that the book's advertisement was shown in the results of the first query and not in those of the second query.

Table 3.4: Overview page of Webometrics Analyst results for the two most visible titles after adding their subtitles to the search queries.

Base query	URLs	Hosts	Domains	SLDs	TLDs
"Governing Failure - Provisional Expertise and the Transformation of Global Development Finance" Best	38	34	34	8	7
"Biological Relatives - IVF, Stem Cells and the Future of Kinship" Franklin	222	168	150	21	18

The least present title on the web was “China’s iGeneration”, which was found on only seven web resources. Because not all of the authors’ surnames were present on most of the web resources, all author surnames except that of the first author were removed from the query. With this modified query, the search returned 169 pages. The tiles in Table 3.5 are arranged from the greatest to the smallest number of unique domains mentioning the title using the optimised search term strategy.

Table 3.5: Final overview page of Webometrics Analyst results for the 28 titles arranged according to the number of domains in which they were present.

Base query	URLs	Hosts	Domains	SLDs	TLDs
"Biological Relatives - IVF, Stem Cells and the Future of Kinship" Franklin	222	168	150	21	18
"Constructing Muslims in France" Fredette	162	115	108	22	21
"Law, Liberty, And The Pursuit Of Terrorism" Douglas	163	120	107	22	20
"China's iGeneration" Johnson	169	121	106	23	21
"Ever Faithful" Sartorious	165	122	105	18	17
"My Voice Is My Weapon" McDonald	150	121	103	18	18
"The Ethics of Armed Conflict" Lango	142	102	95	26	24
"The Myth of Piers Plowman" Warner	138	100	93	26	23
"Composing the Party Line" Tompkins	133	97	90	17	16
"Human Rights and Democracy" Landman	127	99	90	14	13
"Oaths and Swearing in Ancient Greece" Sommerstein Torrance	125	93	88	24	23
"The Emergence of Irish Gothic Fiction" Killeen	131	95	84	18	
"Fighting for a Living" Zürcher	117	93	83	21	19
"Beastly Journeys" Youngs	115	86	79	21	18
"On Global Citizenship" Tully	110	83	79	17	15
"Partisan Gerrymandering and the Construction of American Democracy" Engstrom	109	84	78	22	21
"Passionate Amateurs" Ridout	112	82	76	16	15

Base query	URLs	Hosts	Domains	SLDs	TLDs
"Networks and Institutions in Europe's Emerging Markets" Schoenman	100	75	72	21	19
"Understanding the Global Energy Crisis" Coyle Simmons	116	83	70	19	17
"Aging Gracefully in the Renaissance" Skenazi	93	69	67	23	22
"Rhetorics of Belonging" Bernard	92	68	62	15	12
"Making and Unmaking in Early Modern English Drama" Porter	86	65	62	16	15
"Thinking and Killing" Segev	85	66	60	18	17
"The World Jewish Congress During The Holocaust" Segev	68	46	45	14	13
"In Search of the Amazon" Garfield	48	41	39	10	10
"Electronic Iran" Akhavan	37	35	35	12	11
"Governing Failure - Provisional Expertise and the Transformation of Global Development Finance" Best	38	34	34	8	7
"Verse and Transmutation" Timmermann	37	29	28	7	7

Table 3.5 shows that the most visible title, that is, the title that occurred in the largest number of domains, is “Biological Relatives: IVF, Stem Cells, and the Future of Kinship", which discusses what in vitro fertilization (IVF) means for society. This title name was present in 150 domains. On the other hand, the least visible title is “Verse and Transmutation” by Anke Timmerman, which identifies and investigates a corpus of 21 anonymous Middle English recipes for the philosopher’s stone dating from the 15th century. Likely because of the academic nature of this subject, which may be of interest to a smaller audience, this title name was present in only 28 domains.

In the following section, a content analysis of these two titles is conducted in order to determine who is mentioning these titles. This analysis makes it possible to obtain a more detailed picture of what contributes to high and low visibility.

3.4.5.2 Content analysis

In this section, the URLs in which the 28 title names were present were analysed according to domain types. A classification was done according to organisation represented (e.g., universities, the press, companies, government) or as individual pages (e.g., blogs, personal home pages, social network profiles). An informal

content analysis was conducted, which relied on one person using their personal judgment or intuition of categories. For this reason, a proportional number of each title's results was sampled. Since there were more than 3,200 web resource results for the 28 titles, every tenth matching resource was analysed in order to obtain a general overview of the results. In this way, an appropriate distribution of results was obtained without introducing any systematic bias.

The web resource in which these titles are most commonly present are university site resources, which account for 20% of the total web resources. These include seminar or conference pages, library collection pages, or authors' pages in the university web directory. University sites are followed by bookseller pages, such as Amazon and Barnes and Noble, and other referrals. Around 16% of the results consisted of scam pages. Most of these had similar designs. The sites ask users to register for free to download the titles. Once the user enters their email address, most of the sites ask the user to enter their credit card details on the following page. Although no registration was done by entering credit card information to access the titles' content, these pages were checked with the help of sites such as https://www.scamadviser.com and http://scamanalyze.com/. If the site was categorised as risky, it was recorded as a scam site.

Journal sites constituted more than 5% of the web resources of the sample, where articles cite the titles or review them. This was followed by digital libraries, including HathiTrust, archive.org, and Scribd. The web resource content types are shown in Figure 3.1.

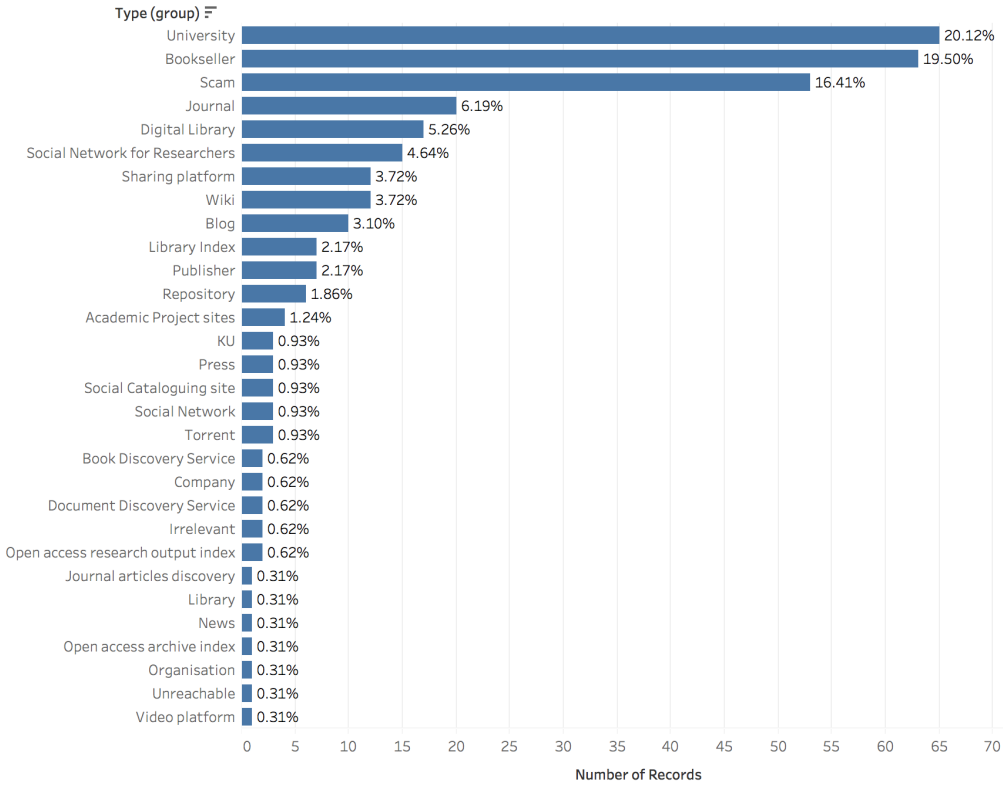

Figure 3.1: Content types for the sample of web resources in which the 28 titles with their authors are present. University-related and bookseller sites constitute the main web resource types.

As discussed in the previous section, all the web resources mentioning the two most visible titles and the least visible title were analysed. For the "Biological Relatives" title, which was present in 222 web resources, seven web resources were found in the Cambridge University domain, where the author of the title, Sarah Franklin, is Professor in the Department of Sociology. This was followed by web resources from Duke University Press; *Science* magazine; the document hosting and sharing platform later converted to e-book subscription service Scribd; the library management system Talis; and the academic publishing company Wiley. A total of nine pages were also found in the American, Canadian, British, and Italian Amazon and Barnes and Noble domains. Scam pages, most of which have short domain names consisting of four letters and belong to German .de TLD sites, such as abdb.de, ajkp.de, avkp.de, or cvee.de, represented 32% of the search results. The title was present in only three pages each from the Blogspot and Wordpress blog sites and from academia.edu. Approximately 14% of the web resources are from university domains. This figure is low compared to the 20% share of university domains for the entire KU pilot collection. There was one page from Wikipedia,

where no link was given to the PDF of the title. Although news sites like BBC Radio or *Science* magazine cite the books on their web pages, they do not provide links to the PDF of the title, but usually present the title's ISBN and price. This does not inform readers that the title is freely accessible online.

The least visible title, "Verse and Transmutation", was present in only 28 domains. The most common category of pages on which the work was present were repositories and publishing or annotation platforms, which made up 13 pages (35% of the total). The second- and third-most common web resources in which the title was present were from booksellers and university sites, with five web resources each. Four web resources were from Brill, the publisher's site, and its free online access site BrillOnline. There were only three web resources each from journal sites and blog sites, and two web resources from the social networking sites Pinterest and ResearchGate.

3.4.5.3 Geographic analysis of TLDs

In this section, the geographic origins of the web resources mentioning each title are identified. First, the domain names of the web resources were extracted, and then TLDs were extracted from these domain names. One challenge is the prevalence of generic TLDs, such as .com, .org, .net, or .info, which are not nation-specific.

Most of the 3,238 web resources in which the 28 titles were present (2,334, 72.1%) were from domains in English-speaking countries (the United States, United Kingdom, Canada, and Australia). To examine the domains in which these titles were present, country-independent TLD's, including .com, .org, .net, and .info, were filtered out. This filtering left 1,084 domains. However, this was done at the expense of removing domains from the United States, excluding academic ones. It was found that the domains in which the title names were most frequently present (355, 32.7%) were .gov, and .edu domains (excluding academia.edu), which point to academic institutions in the United States. With the exception of a few registered institutions outside of the US, .edu domains are used by academic institutions in the US (Cooper, & Postel, 1993).

The U.S. resources were followed by web resources from the UK (264), Germany (100), Italy (65), Canada (57), and Australia (46). Most of the German TLD .de sites

were scam sites, and most of the Italian TLD .it sites were from unglue.it, which is a platform for the distribution and storage of free ebooks. Figure 3.2 shows the top TLDs for web resources in the Bing Search results for the 28 titles after country-independent TLDs were removed.

Second-level domains were used to classify the domains' entity types. Certain domains were classified as academic, including .edu, .ac.uk, .ac.jp, .edu.au, .edu.cn, and .ac.il. Domains such as .com, .co.uk, .com.au, and .co.jp were classified as commercial. While the entity type of domains belonging to certain TLDs, including Germany, Italy, Canada, and France, could not be definitively classified, it was clear from manual inspection that the top three types of entities were commercial, academic, and organisational. The distribution is shown in Figure 3.3.

Academic domains accounted for less than 21% of the domains in which the monograph titles were present. Even when .com, .org, and .net domains are set aside and uncategorized domains are classified as academic, academic websites make up at most 34% of the total domains. In other words, academic domains constitute one third or less of the total number of domain types in which these title names were present.

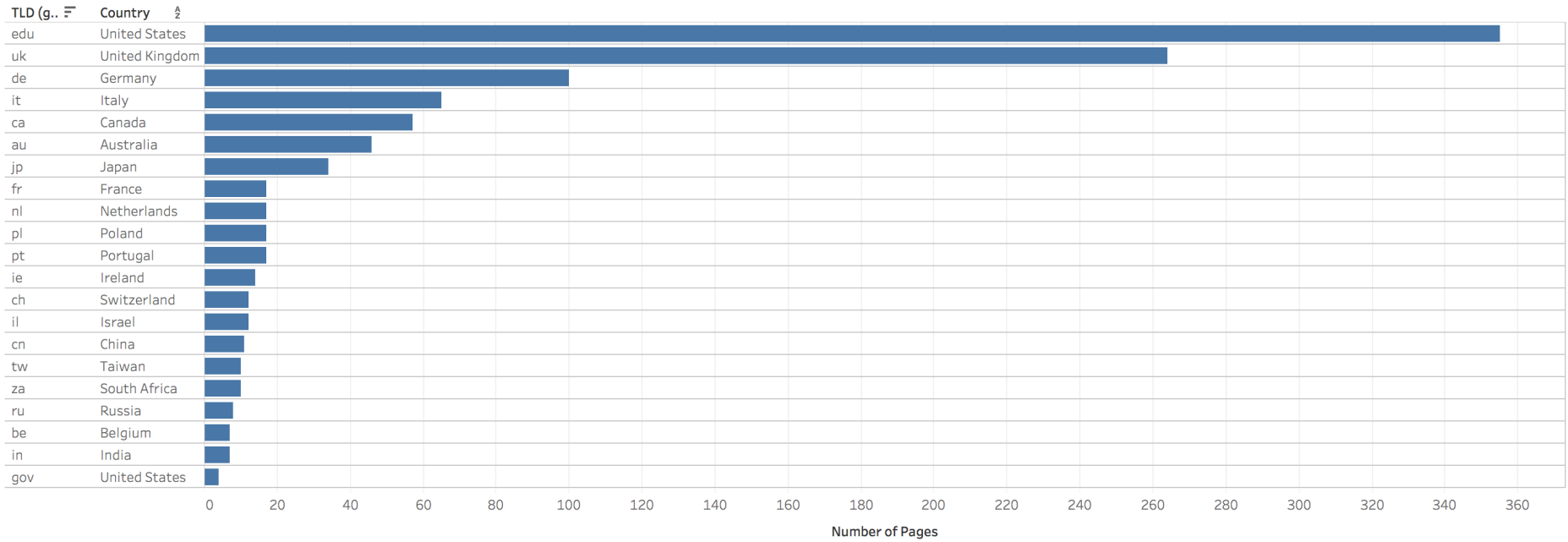

Figure 3.2: Top-level domains with the highest number of URLs in the Bing Search results.

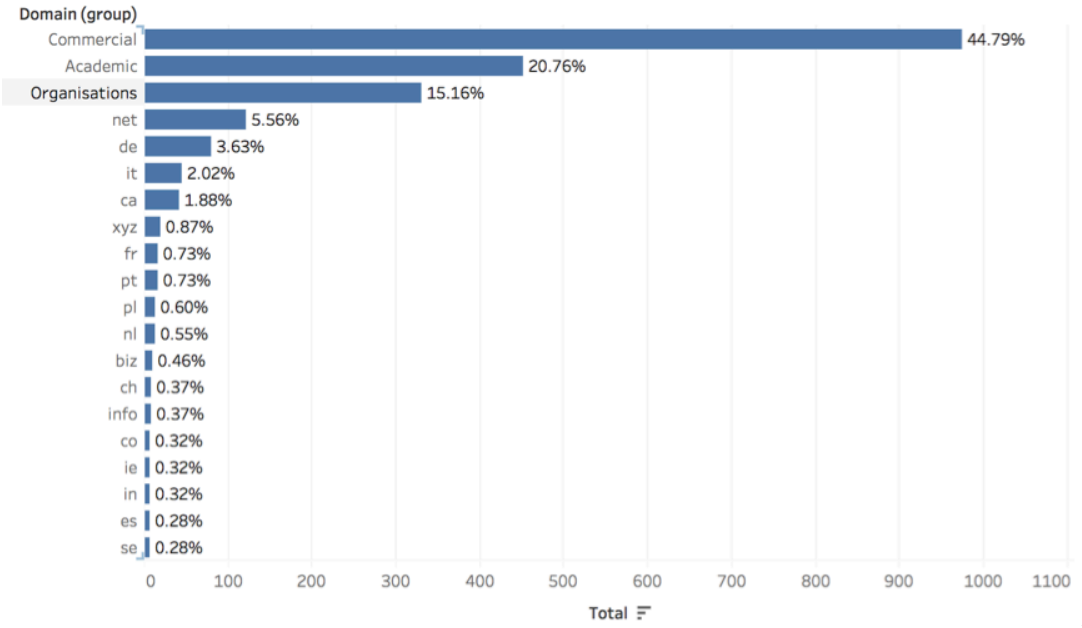

Figure 3.3: Distribution of organization types of domains in which the 28 title names are present.

3.4.5.4 Repository presence

To determine the visibility of the repositories for the 28 KU titles on the web, the URL of all the links in all the web resources where the title names were present were collected and investigated. From the 3,190 web resources (shown in Table 3.5) that contained the 28 title names with their author names, 429,649 links were scraped. The links' domains were filtered using the strings 'hathitrust.org' and 'oapen.org'.

The number of web resources containing links to OAPEN and HathiTrust were 186 and 95, respectively. Among these web resources, 27 title names were present in web resources with links to OAPEN and 26 title names were present in resources with links to HathiTrust. The domain names of the URL results obtained in Section 3.4.5.3 were also examined. Among 3,190 URLs, 55 domains were identified as belonging to OAPEN and 37 domains as belonging to HathiTrust. The URLs belonging to OAPEN covered 28 titles, while the URLs belonging to HathiTrust only covered 19 titles.

These results show that the OAPEN repository is much more visible on the web than the HathiTrust repository. The high visibility of the OAPEN repository is due to the high discoverability of its titles. The fact that the OAPEN repository is registered to the DOAB and library catalogues and is indexed by search engines makes the 28 titles in this repository more visible to other sites in which these titles are present.

The difference in visibility between OAPEN and HathiTrust is mainly due to the difference between these repositories' purposes. Although both repositories store books, OAPEN hosts OA academic books, mainly in the HSS area, and it assists publishers and libraries in disseminating these books. HathiTrust, in contrast, aims to preserve a record of human knowledge, and therefore contains more than 14 million volumes, of which OA books constitute a single part.

The following section compares the access to these repositories.

3.4.6 Discussion

The visibility analysis found that links to the 28 KU titles were rare in web resources. When links did occur, the majority referred to the publishers' pages. Approximately half of the publishers did not provide a link to the book's OA PDF version. This may be because publishers are concerned that providing links to free versions may reduce book sales. The lack of links to OA PDF versions of these titles may be explained with reference to traditional publishing models' approach to OA monographs (Bonn, 2015). Traditionally, publishers have assumed that saleable objects will result in revenues that can be used to cover costs and generate profits to expand their business. This creates an incentive to market these objects in order to bring them to the attention of the purchasing public, which includes both individuals and entities such as libraries. From this perspective, if there is nothing to sell, there is little point in marketing. However, this approach has started to change. In fact, a study conducted by OAPEN-NL focusing on the period between 2011 and 2012 found no evidence that making monographs OA reduces print sales. On the other hand, their report states that online access to OA books increased considerably, as did online discovery of these books (Ferwerda, Snijder, & Adema, 2013). Some publishers, such as SpringerOpen have started market to authors and funders by

publicising their ebooks' access statistics and altmetrics attention scores (The Digital Methods Initiative, n.d.).

3.5 Accessing monographs

Webometrics can be used to reveal the visibility of the 28 KU titles by indicating where and by which sites they were mentioned. This section focuses on access statistics, which are used to map how, when, where, and by whom titles are accessed. Two types of access data were used, namely download counts and web traffic statistics.

Because the KU pilot collection titles are hosted on the OAPEN repository, this section mainly analyses OAPEN's access statistics. Download count reports and web traffic statistics for each title for the period of March 2014 to June 2017 are investigated. In addition to OAPEN's access statistics, HathiTrust web traffic statistics are used as well.

This section starts by reviewing the issues on access statistics relating to OA monographs. It then discusses two different types of monthly download counts provided by OAPEN, namely country-based and IP address-based download counts. IP address-based download counts are helpful to identify downloads coming from institutions and also make it possible to geocode users' locations, which enables access from specific regions to be examined. The section then discusses web traffic statistics to analyse where users are coming from and identify channels that are effective in directing traffic to the repository.

3.5.1 Access reports

Every repository manager is trying to optimise and improve their repository and also to demonstrate the repository's value to authors and publishers. In order to support their operations, repositories collect different kinds of access metrics in different ways, which some then make available as services to their members. However, the issues relating to the combination and comparison of access statistics of OA monographs differ from those of journals. These issues generally arise because different types of platforms have different types of services and different types of business models, which entails the use of different techniques for gathering access statistics.

Firstly, OA monographs can be hosted on many different platforms, including publishers' platforms, as in the case of UCL Press and Ubiquity Press; large-scale repositories such as OAPEN, OpenEdition, and HathiTrust; digital libraries such as JSTOR or The Internet Archive; and institutional repositories. Some of these platforms make their monographs accessible to readers in different ways, according to their business models. OAPEN offers free downloadable PDFs of monographs. Its participants and the Netherlands Organisation for Scientific Research (NWO) provide an annual subsidy, and the remainder of its income comes from services and projects (OAPEN, n.d.). OpenEdition, on the other hand, uses a "freemium" model, which provides free access to books in on-screen viewable HTML format, but which also allows members (mainly libraries) to pay to receive additional PDF and ePUB file access, as well as various other services. Non-members can only see the HTML versions of the books for free (OpenEdition, n.d.).

There are also different types of repository software on the market, such as ePrints or DSpace, which come with a variety of add-ons for access tracking. Various third-party solutions are also used, including Google Analytics, Piwik, and Adobe Analytics. All of these process raw access data in different ways, so there is a lack of agreed-upon standards to measure access across repositories (Needham & Stone, 2012). Thus, it becomes difficult to compare and benchmark access statistics.

Gaming in access statistics is another challenge. Since the late 1990s, website owners have tried to attract more users using advertisements and search engine optimisation techniques. Since page-views and clicks are important, spammers send high volumes of nonsense emails, display irrelevant web pages to bring people to their websites, or use techniques to increase web page views. The simplest gaming method is for website owners to increase access statistics by repeatedly loading a page. According to Zeifman (2015), in websites that see fewer than 10,000 visitors per day, it is estimated that less than 30% of online traffic is human-initiated.

Academic repositories face similar issues: if downloads and page view statistics are used as metrics for promotion and funding, there is an incentive for researchers to attempt to game them (William, 2017). The Council of Australian University Librarians (CAUL) recommends that in order to keep statistics useful, they should not be tied to rewards (CAUL, 2017).

One of the international efforts to overcome these problems is the Counting Online Usage of Networked Electronic Resources (COUNTER) project (https://www.projectcounter.org/). COUNTER is a code of practice for compiling online usage statistics for electronic resources. The COUNTER code of practice includes rules such as removing robot entries and double clicks. This code of practice is intended to help publishers and vendors support their library customers and provide statistics comparable to those of their competitors in a consistent and credible way. This is especially important in the context of subscription content, which libraries may pay for on a per-use basis. OAPEN, which was originally one of the two main repositories for the KU pilot collection, also cooperates with IRUS-UK, which provides COUNTER-compliant reports. IRUS-UK enables institutional repositories to provide and share statistics based on the COUNTER standard. It provides a nation-wide view of UK repository usage to benefit organisations such as Jisc, which is a British not-for-profit organisation whose role is to provide digital services and solutions for higher education and research (Jisc, n.d.). IRUS-UK offers opportunities for benchmarking and acts as an intermediary between UK repositories and other agencies (IRUS-UK, n.d.). OAPEN provides KU with usage reports for their PDF downloads that have been prepared using the COUNTER methodology. IRUS-UK provides two types of COUNTER book reports for the OAPEN repository: Book Report 1 (BR1), which indicates the number of successful title requests by month and title; and Book Report 2 (BR2), which provides more granularity by providing the number of successful section requests by month and title for their web-viewed titles. In this study, which focuses on KU titles, BR1 reports were used. These reports provide access statistics by country or by IP address.

The other repository investigated in detail in this thesis is HathiTrust. HathiTrust makes available HTML versions of books and relies on Google Analytics to collect page view numbers for its books. The OAPEN repository also uses Google Analytics to collect data on website page views. However, unlike HathiTrust, OAPEN does not provide HTML versions of the KU titles. Instead, OAPEN uses their website to present these titles and provides a link on the title's presentation page to the PDF download in their repository.

Since web traffic statistics represent different types of access, it is not possible to combine them to obtain an aggregate access figure. However, these distinct types of access are helpful to obtain an idea of how these titles are read, since Google Analytics is able to track user activities on websites.

3.5.1.1 Benchmarking monographs

Comparing a title's access statistics with those of other titles in the same repository can be helpful in order to obtain a rough estimate of a title's performance. However, for benchmarking, a comparison with other books with similar attributes is needed. First, it is necessary to extract books in the same language, since the same book in two different languages, such as English and Dutch, will have different numbers of potential readers. Comparing an English HSS title with other English titles is useful to determine how much attention the HSS monograph has received.

However, comparing an English title only to an average English-language monograph is not particularly informative. More shared attributes are needed to make an appropriate comparison, for example, relating to the field within which the books are classified. As Schubert and Braun (1996) argue, "mere publication or citation counts are completely inadequate measures of scientific merit; they can be used for evaluative purposes only after proper standardization or normalisation". On the basis of Schubert and Braun's idea, monographs were evaluated according to their field by using field normalisation. As Ioannidis and colleagues state, normalisation can be seen as a process of benchmarking that is needed to enhance comparability across diverse scientists, fields, papers, time periods, and so forth (Ioannidis, Boyack, & Wouters, 2016). Using field normalisation, monographs were benchmarked against other monographs within the same subject fields. In order to do so, books in the OAPEN repository were filtered according to their subject using the repository title catalogue metafile which contains the subject, language, upload dates, publication dates for each title. Books on the repository were also filtered so that their publication and upload dates to the repository were within the same period as those of the KU pilot collection titles with which they were compared. After books in comparison group were identified for each title, access figures were downloaded. Subsequently, the comparison group's access average was calculated and compared with the title. These books were also filtered so that their publication

and upload dates to the repository were within the same period as those of the KU pilot collection titles with which they were compared.

A high access figure does not indicate whether a particular monograph is higher in quality or more informative than other monographs; it merely shows that it is accessed more frequently. In the following sections, repository download counts and web traffic statistics are discussed in relation to how useful information can be extracted from these figures to obtain a more detailed overview of monograph access.

3.5.2 Download counts

In this study, download counts are based on COUNTER-compliant reports provided by OAPEN. Two types of COUNTER-compliant reports are used, namely country-based and IP address-based monthly download counts for each title.

3.5.2.1 Country access

Using COUNTER BR1 reports, it is possible to track titles' access by country on a monthly basis. This is especially important in the case of OA publishing, which removes financial barriers and allows unrestricted access to scholarly information for people across the globe. Figure 3.4 shows the 86,202 downloads of the 28 KU pilot collection titles that occurred between March 2014 and June 2017 on a map. The highest downloads (shown in dark blue) were from the United States followed by the United Kingdom.

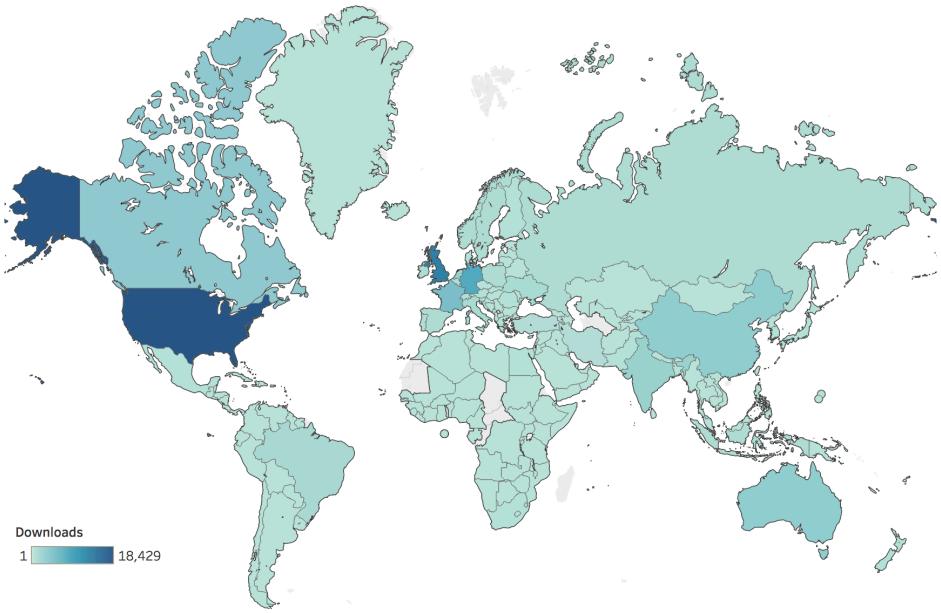

Figure 3.4: Global downloads of KU pilot collection titles from OAPEN from March 2014 to June 2017.

It was found that the number of downloads from the geographic locations mentioned in the books' subject metadata was higher compared to the number of downloads for other titles for the same locations. This is important and worth mentioning, since OA reduces the financial barriers to accessing scholarly content, which is important for developing countries. However, in this study, there was a lack of evidence linking a book's high access from a specific geographic location to its subject because of the title's OA attribute.

3.5.3 Institutional access

The country-based access report provides approximate information about access to books. To obtain more detailed information on this matter, I investigated institutional access for KU as part of the KU Research team. This involved preparing an institutional access dashboard, which was done by extracting institutional IP address blocks from the COUNTER IP address access reports provided by OAPEN. I then analysed access for each institution.

Libraries need institutional access information to inform their funders of the usage within their institutions of book titles the funders have chosen to support via KU.

Institutional access data is also helpful for this thesis, as it sheds light on libraries' role in spreading knowledge in the HSS field.

One limitation of IP-address-specific institutional access is that this information excludes access by members of a community that occurs outside of the university's IP range (e.g., a university researcher accessing the content from home or from a mobile device).

In this study, to obtain more fine-grained information and a more complete overview of where the 28 KU titles were read outside of university campuses and libraries, downloads were geolocated, beginning with IP address access reports. The OA attribute enables researchers to access these titles outside of university campuses or libraries as well.

Geolocating access from outside of university campuses can also provide an approximate indication of how OA enables the downloading of content by groups outside of the university who may not have engaged with the content if it had remained behind a paywall.

To determine the geographic locations of IP addresses, the addresses were geolocated to obtain their latitude and longitude. This was done by means of the 'rgeolocate' R package, using the Maxmind GeoLiteIP service. After finding the IP addresses' latitudes and longitudes, the geolocations were reverse geocoded using the 'RGA' R package and the Google Maps Geocoding API in order to obtain city names. Google Maps reverse geolocates locations according to their administration levels. According to Google Maps, administrative area level 1 indicates a first-order civil entity below the country level. For example, within the United States or Australia, these administrative levels are states. In the UK, these represent four countries (England, Northern Ireland, Scotland, and Wales), and in France, 13 metropolitan regions. Administrative area level 2 indicates second-order civil entities below the country level, which would be counties for the United States, departments for France, and cities for Australia.

There are also challenges in associating IP addresses with geolocations. Firstly, these IP address geolocations are not 100% accurate. IP addresses may be associated with incorrect locations, such as incorrect postal codes, cities, or suburbs within a

metropolitan area. They can also be associated with very broad geographic areas, such as large cities or states. In fact, many addresses are associated only with a city, not with a street address or latitude/longitude location.

Another issue is that some IP addresses are not in the database and therefore cannot be mapped. In addition, some users connect to monograph repositories through proxy servers. These proxy servers are computer systems or applications that act as intermediaries for requests from clients seeking resources from other servers. However, in the case of this study, it is not particularly important if users connect through library proxies, as this just indicates that they are physically off campus, likely outside of working hours, which does not matter for our purposes. One of the challenges faced during this process is changes in the geolocations of IP addresses over time, because of release, allocation, and reallocation. These changes are why companies update GeoIP databases on a monthly basis and charge users to access these databases.

As a result, when using an up-to-date database, the likelihood of successfully locating an instance of access that took place a year ago is reduced. Geocoding access based on IP addresses using a geolocation database belonging to the relevant access period is more accurate. From a practical perspective, the easiest way to accomplish this is by adopting a periodic approach to geolocation in order to avoid the complexities of attempting to geolocate access retrospectively. Starting from January 2017, to increase accuracy, downloads were geolocated on a quarterly basis using an up-to-date database.

Although the accuracy of IP addresses' locations changes over time, the company that provides the database that was used (MaxMind GeoIP2) states that their database is 99.8% accurate on a country level, 90% accurate on a state level in the US, and 81% accurate for cities in the US within a 50-kilometre radius. However, accuracy levels vary for individual countries. Nonetheless, geolocating IP addresses is useful to provide an indication of where books are downloaded from. This is helpful in order to identify downloads in close proximity to universities. Since the content being accessed is OA, it does not need to be downloaded from institutional IP blocks. Readers from universities can also access this content from their homes. State-based access with a 90% accuracy rate is also important, in order to give an

indication of how state-funded libraries are helping their states to provide the titles they have supported.

3.5.3.1 Book Downloads vs. Chapter Downloads

Not all platforms provide whole-book downloads; there are also some platforms, including JSTOR, which make OA books available in the form of chapter-level downloads. JSTOR, a digital library founded in 1995, provides access to academic journal articles, books, and other type of sources in 75 disciplines. In October 2016, JSTOR began providing OA books, initially from four publishers, including University of California Press, University of Michigan Press, UCL Press, and Cornell University Press (Montgomery, Ozaygen, Pinter, & Saunders, 2017). However, comparing whole-book downloads with chapter downloads is akin to comparing apples and oranges. It also does not make sense to divide the number of chapter downloads by the number of chapters in a book in order to obtain a total number of 'book' downloads, because this would change from book to book. In addition, not all readers read all the chapters of a book, and they do not need to read from the beginning of a book. Sometimes, scholars only read the chapter they need for their research.

There are also dangers in attempting to compare chapter downloads from different platforms. Individual platforms are built using different architectures, and they provide services in different ways. For example, when searching in library catalogues, search engines, or on the JSTOR site, a list of book chapters from the JSTOR site may be encountered. Since there are no abstracts for book chapters on the resulting page, the user has to download the chapters that appear in the results list in order to determine whether they match their needs. It is likely that some of the chapters that are downloaded are not read at all. To overcome this problem, JSTOR has begun displaying a snippet of the text where the search terms are used, and they also provide the topic of the book's chapter by calculating the frequency of words in chapters to help users find what they are looking for (JSTOR, 2016).

3.5.4 Web traffic statistics

In addition to providing books as whole books or as chapters in PDF, EPUB or other file formats, some platforms offer them as online web page views. Depending on

their business model, some provide a freemium service, such as selling the PDF or EPUB versions and offering the HTML version of the book for free. HathiTrust, OpenEdition, Open Book Publishers, and OAPEN offer web page views of their books.

Although providing book contents in HTML or image format forces readers to read with their browsers connected to the web, it can provide valuable information about online readers. This information on access can be helpful to understand how readers interact with book content. It makes it possible to identify the sections of books that are read most and the time users spend on different sections or types of content.

In order to gather data on web page access and user interaction with their sites, publishers and repositories rely on web analytics software. There are two methods of collecting access data from digital repositories using web analytics software. The first is page tagging and the second is via the analysis of log files.

Page-tagging analytics, such as Google Analytics or Adobe Marketing Cloud, are offered as Software as a Service (SaaS) from the vendor's website. These are usually based on a script, which is placed in each HTML page of the website in question to track access.

Each time a web page is displayed, it triggers a signal from the tracking code to the software and the software registers these visits. Each visit record can include information such as the user's geographical location, the type of operating system they are using, the site they were directed from, how much time elapses between each click, and how many users navigate away from the site after viewing only one page. This information is helpful in understanding users' interaction with a website and in identifying effective channels in the dissemination of information.

The other method of collecting access data is based on software that analyses log files residing on the server side, where all website events are recorded. Event analysis software usually prepares reports in a file or displays them on a web page. These log file analytics can also be included in repository software packages such as DSpace or ePrints.

Page-tagging analytics and log file methods have different advantages. For example, page-tagging analytics can be used to determine how users interact with a repository,

such as the path they follow when navigating the site. On the other hand, log files can provide fine-grained information about repository access, but pose a considerable risk in terms of under- or overcounting visits, downloads, and page views. Obrien and others (2016) analysed these two methods and found that using page-tagging analytics runs the risk of undercounting non-HTML file downloads, particularly when users are referred directly to the file from an external source. For example, if a user is referred directly to the PDF file, no page tag is called for, and thus no activity log on the server is triggered. With the use of web logs, however, there is a significant risk of overcounting the number of downloads or page views, because it is not always possible to filter bots, crawlers, and scrapers.

The standard configuration of Google Analytics provides only page view statistics. In order to track non-HTML content downloads, an additional configuration called "event tracking" must be used (Bragg et al., 2016). For example, in order to count downloads, Google Tag Manager must be employed.

3.5.4.1 Alternatives to Google Analytics

Google Analytics is not the only web analytics service on the market. There are tens of alternatives, which can be installed as free or proprietary software or used as cloud services. Google Analytics is offered as a service using cookies. A significant proportion of publishers and repositories use the free (standard) Google Analytics service. According to W3Techs.com, a website that monitors the global market share of web technologies, Google Analytics is the most-used web traffic analysis tool on the market, with a market share larger than 80% (W3Techs, 2017). OBrien and colleagues also found that Google Analytics tracking code was used in over 80% of the 263 academic libraries they surveyed (OBrien et al., 2016).

The EU Cookie law, enacted on May 26, 2012, states that if a site uses cookies or tracking technologies for a non-essential function, it must:

- Tell users that tracking technologies are used
- Explain the reasons for using these technologies
- Obtain the user's consent prior to tracking them and allow them to withdraw permission at any time (Buckler, 2012).

This European privacy law places server-side software packages such as Piwik in an advantageous position. Since these self-hosted software packages do not require cookies, sites are able to comply with the EU Cookie Law without obtaining users' consent to use cookies. There are also more expensive enterprise solutions favoured by large publishers or archives, such as Adobe Analytics, IBM Analytics, and Google Analytics Premium. Some of these analytics solutions focus primarily on the performance of sites and pages; and some focus primarily on where visitors come from and what they are doing on a site, or they have other features focusing on the site's ecommerce aspects. The present study is based on the most widely used free analytics service, namely Google Analytics. However, the same principles apply for other analytics. The following sections report on the use of some of the Google Analytics features in order to understand access to OA books.

3.5.4.2 Book web pages

In addition to tracking book content usage, web analytics software can also be used to track traffic on other types of book web pages. In general, when users access a monograph, they are directed to the content by one of three types of web page. The first type are publishers' sites which present the book. On these sites, the book is likely to be sold in other digital formats or in print. Sales may occur via companies such as Amazon.com, and Google Play or the book may be sold directly from the publisher's domain. The publisher's site may also direct users to repositories where they can download a PDF of the monograph or read the content online. These web pages on publishers' sites are categorised as "webshops" in the OAPEN repository.

The second type of web page directing users to content are directory web pages in which OA books are indexed, such as the DOAB. These indexes are contained in library catalogues and act as a direct discovery service for OA monographs. They provide a searchable index for users and provide a link to the full texts of books.

The third type of page are web pages on repositories that also present books, usually on one page, with links to the file(s) or to the web pages that display the content.

Since these three types of web page direct users to the repositories, web analytics software packages provide valuable information about users' behaviour when

accessing these pages. This study uses the web traffic data of book presentation pages on the OAPEN repository.

3.5.4.3 Web analytics metrics

Web analytics services and software provide different types of metrics with which to capture user behaviour on a site. This section discusses five different metrics: page views, unique page views, sessions, bounce rate and traffic sources.

3.5.4.3.1 Page views, unique page views and sessions

Google Analytics uses different terms to define the number of times a page has been viewed. These are "page views", "unique page views", and "sessions". "Page views" are the total number of pages that have been viewed. Repeated views of a single page are also counted. This means that if a user navigates to a different page and then returns to the same page, a second page view is recorded. Unique page views, on the other hand, refer to the number of sessions during which the specified page was viewed one or more times. Since the present study is concerned with each session interaction, because it generally represents one visit, unique page views are used to count the number of separate visits to a page. This is also in accordance with the COUNTER methodology for quantifying usage, which does not count repeated downloads. Use of unique page views helps to discard artifacts such as users increasing a page's page views by continuously reloading the page. An indication of book's access can be obtained by comparing the unique page views of a book presentation web page on a repository with the book's downloads. This may give an indication of how many of these sessions result in book downloads.

In this study, rather than number of users, number of sessions was used. This is because in Google Analytics, users are tracked using cookies that are downloaded to their computer's web browser. Thus, when a user connects to the site with the same computer but using another web browser, they are counted as a new user. On the other hand, if a user employs a web browser on a public computer to connect to the site, they will be counted as the same user who connected to the site previously from the same computer. For this reason, it is more relevant to count the number of visits instead of the number of users. Since the number of visits (sessions) to each book page is the same as the number of unique page views for that book, which is equal to

one page of presentation for each book in OAPEN, the number of unique page views per book was compared to each book's respective PDF downloads.

3.5.4.3.2 Bounce rates

Google Analytics provides metrics regarding a site's bounce rate, which is the percentage of visitors entering or landing on a website and leaving without continuing to another page on the site. This metric is generally used by site owners to understand whether the content on a website is what users are looking for. For example, users can land on a website and leave it because it may not be what they are looking for. However, bounce rates can in some cases be interpreted differently, depending on the site's purpose and architecture. Bounce rate metrics may provide information about users who are only coming to download a specific title on a repository or indexing site rather than browsing the site to see the other titles that are displayed. For example, in the case of the OAPEN repository or the UCL Press website, users may land on a title's web page, click on the "download PDF" button, and be forwarded to a PDF download site, which may be on another domain. In this case, the site's bounce rate would increase, and it would probably also indicate that this user is the same user counted as downloading the title (although this needs to be verified with a comparison between the unique page views and the downloads for each book). However, contrary to this example, in the case of The Internet Archive, PDF files reside under the domain name of the book's web page, and so the bounce rate for this site would be low.

To track users who click on a link to leave a site, Google has created outbound link tracker code, which is not often used by website administrators ("Track outbound links", n.d.). In the present study, titles' downloads and their unique page views on OAPEN are collected using different methods and displayed with different granularities. For this reason, these accesses cannot be accurately compared to interpret a page's bounce rate.

3.5.4.3.3 Traffic Sources

For publishers, it is also useful to understand where readers are coming to their sites from. Google Analytics provides the origin of the traffic visiting the site, which assists publishers in identifying the channels that are effective for disseminating

books or for the discovery of monographs. Using traffic source information, publishers can determine which sites drive more traffic, identify marketing or distribution channels to which they should pay more attention, or obtain information about how they can improve their discovery and reach. For example, KU Research, as part of their UCL Press usage project, found that one of the biggest traffic-driving channels to the UCL Press website for disseminating their books was emails to listservs (Montgomery, Neylon, Ozaygen, & Leaver, 2018).

To gain more granular information on the sources of website traffic, it is possible to cross-reference traffic instances with marketing events initiated by the publisher. This can inform publishers about the effectiveness of events they have organised for disseminating their books. As part of the UCL Press project, KU Research plotted book downloads against key events. A significant increase in download rates was observed just after the start of a massive open online course (MOOC) program that pertained to a specific book. To obtain a more detailed picture of the access to this specific book, the team analysed the sources of traffic to the book's page. They noticed that the download figures for this title were much higher than the unique page views of the book presentation web page. After checking the web page statistics, they noticed that the site of the MOOC programme in which the book was used was not among the traffic sources. Subsequently, after checking the MOOC program pages, it was noticed that the link provided referred directly to the book's PDF file, instead to the book's page on the UCL Press website. This examination of different types of access data makes it possible to infer where users who download PDF books are most likely to have come from.

The Google Analytics traffic source feature also makes it possible to determine whether book titles occur on websites from the region they describe. This is useful, as it indicates whether local community websites show interest in these titles, which can result in downloads. This would also indicate whether the OA feature of the monographs is benefitting local communities. For example, as part of their UCL Press usage project, KU Research found that some books related to specific regions are read more in these regions.

3.5.4.4 Access and Events Comparison

In the UCL Press usage project, the KU Research team investigated whether marketing and promotion activities had any effect on monograph access. This investigation was useful, as it served to identify events that were efficient in promoting titles. As mentioned in the previous section on traffic sources, KU Research noticed that the UCL MOOC program that used UCL Press titles and some mailing lists UCL Press used to promote their books were considerably more efficient than other events. Using a similar approach, the following sections unpack the reasons behind the spikes in book access for titles in the KU pilot collection. That is, these sections work backwards from access spikes to events, instead of examining how certain events affect access (Montgomery, Neylon, Ozaygen, & Leaver, 2018).

3.5.5 Findings

3.5.5.1 COUNTER-compliant country-specific access

The KU pilot collection was originally hosted on the OAPEN, HathiTrust, and Internet Archive platforms. Since 2017, they have also been uploaded as separate chapters on the JSTOR platform. OAPEN has provided COUNTER-compliant usage reports for 28 titles. Since these titles were uploaded to the repository separately and on different dates between March and September 2014, monthly download averages were obtained for the entire KU collection by dividing each month's aggregated downloads by the number of titles present in the repository in that month.

Figure 3.5 shows the average downloads per month for the KU pilot collection titles. Access decreases during the months of June, July, and August, when it is summer in the Northern Hemisphere.

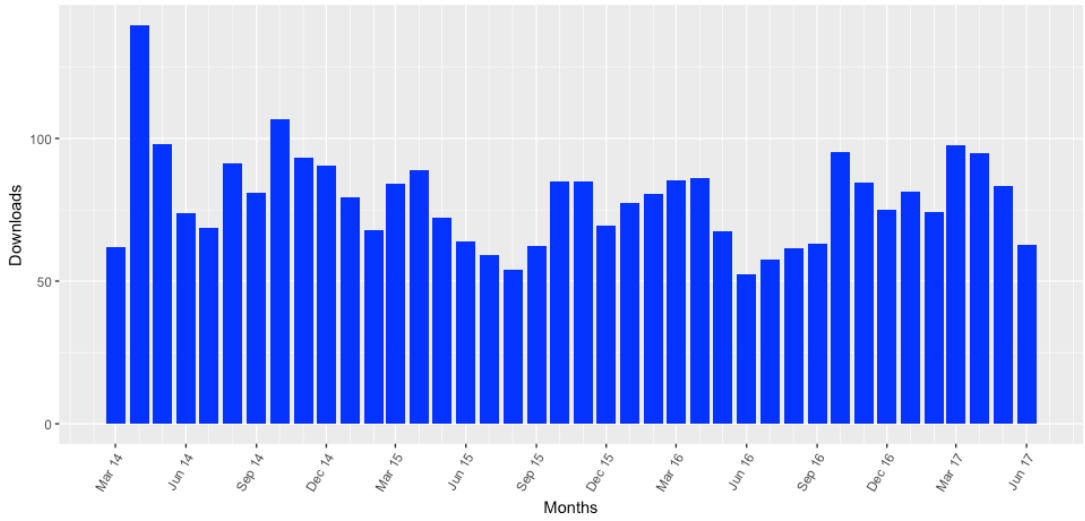

Figure 3.5: The KU collection's aggregated monthly download averages from OAPEN.

Since these titles were uploaded to these platforms at different points in time, average monthly downloads were calculated for each title, as shown in Figure 3.6.

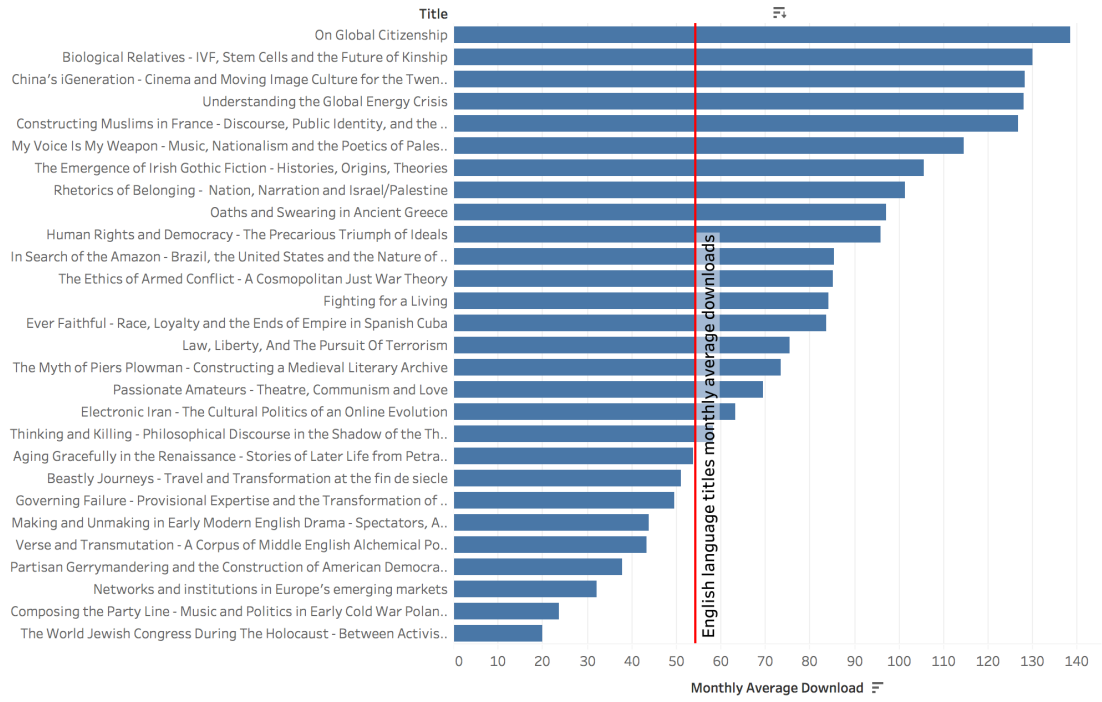

Figure 3.6: KU pilot collection titles' average monthly downloads from OAPEN, including line indicating the "English-language titles' average monthly downloads".

Figure 3.6 shows that the title with the most average downloads per month is "On Global Citizenship" title, with 138.5 average downloads per month. The title with the lowest download rate is "The World Jewish Congress During the Holocaust" title, with 19.8 average downloads per month. These titles' downloads were compared to the average monthly downloads of all the English-language

monographs on OAPEN. To do this, English-language titles were extracted using the repository metafile. Titles that were published after 2010 and uploaded to the repository before March 2014 were included, and 721 titles were obtained. It can be seen from Figure 3.6 that 19 out of 28 titles were downloaded more than the average English-language title.

The monthly distributions of the English-language titles and the 28 KU collection titles' downloads is plotted in Figure 3.7. The average downloads of English-language titles' are lower than the 28 KU collection titles' downloads because they contain a high proportion of low monthly downloads.

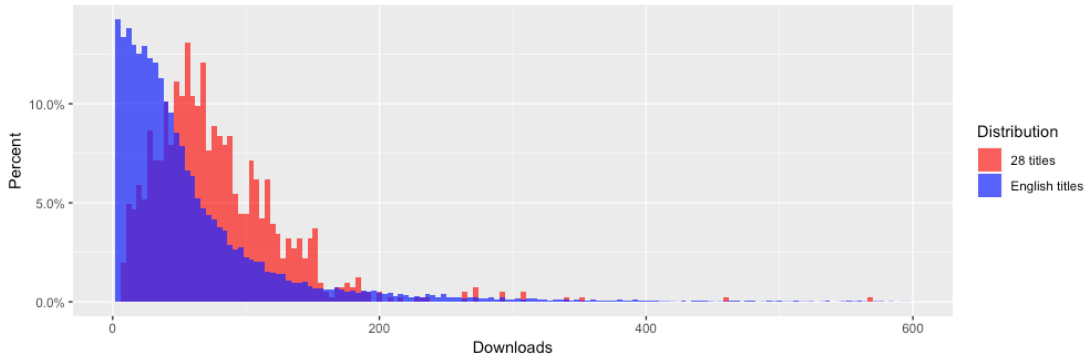

Figure 3.7: Monthly distribution of English-language titles and the 28 KU pilot collection titles

Each title's monthly downloads were also plotted in order to identify common patterns and determine whether there were any unusual download patterns, as shown in Figure 3.8.

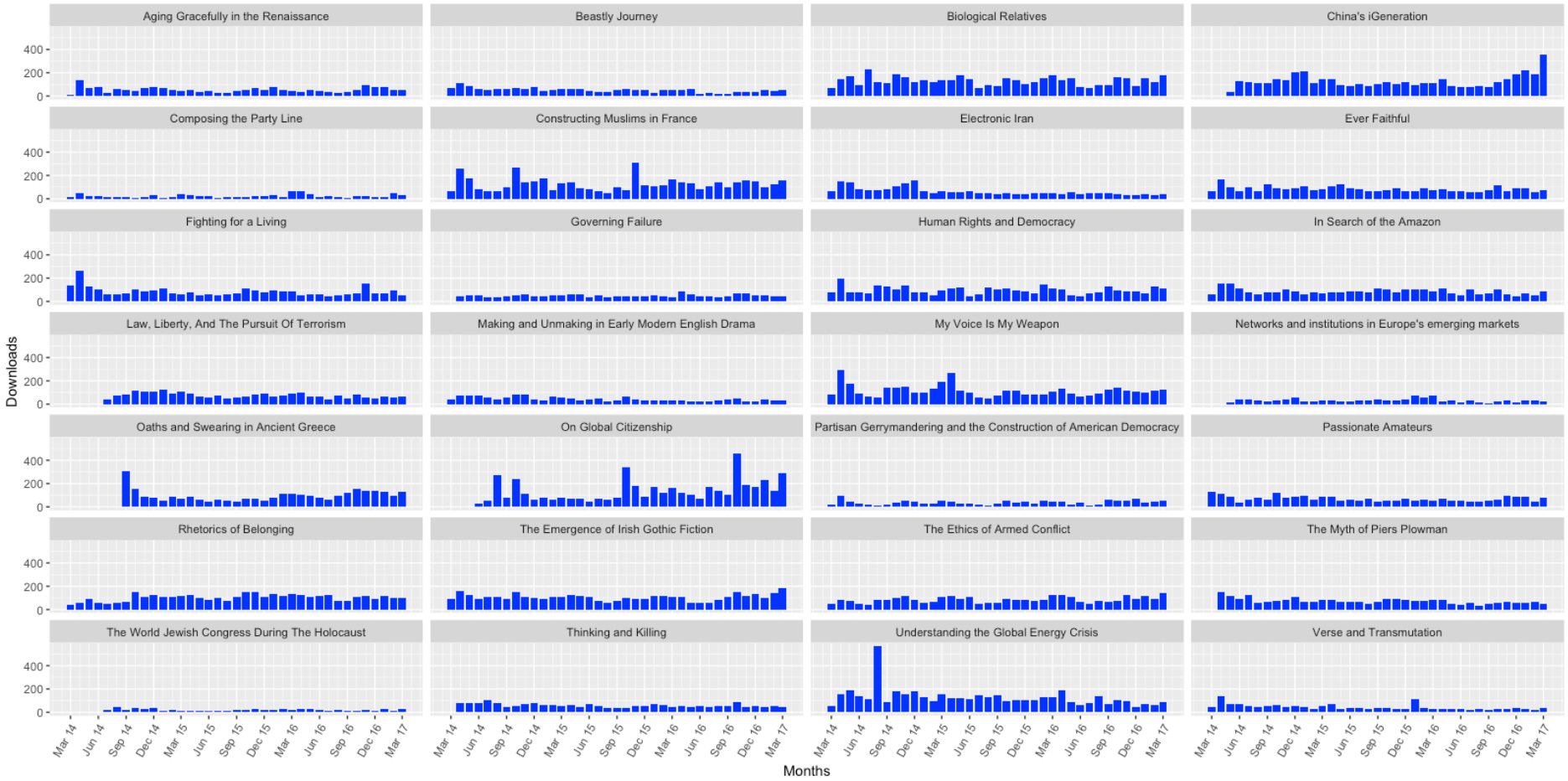

Figure 3.8: KU pilot collection titles' monthly downloads from the OAPEN repository.

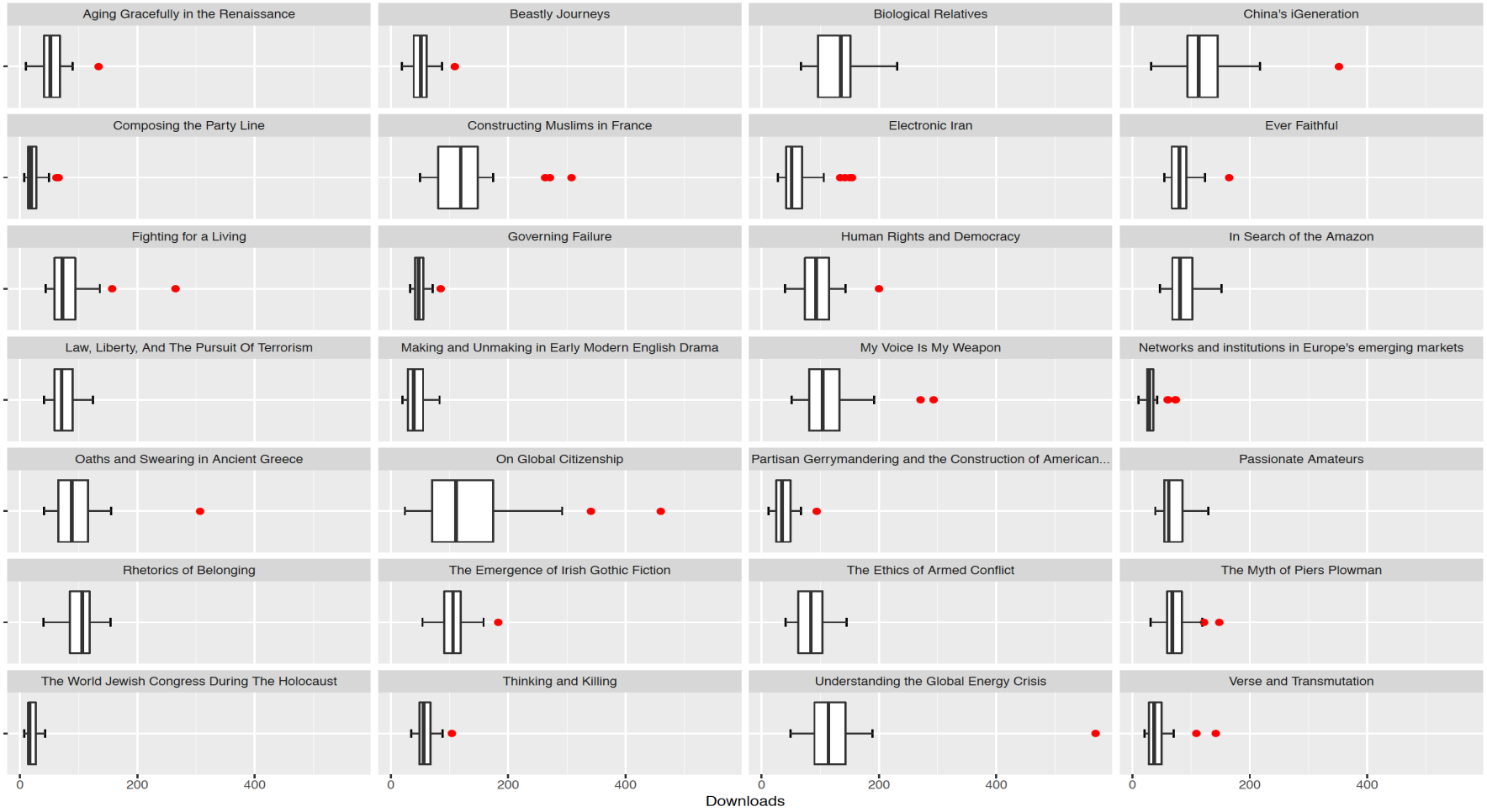

Figure 3.9: Monthly download distribution graph for each title, where outliers are shown as red dots.

In Figure 3.8 no common patterns were observed. However, some peaks were much higher than others. To make sure that some of these peaks were in fact outliers, each title's access was also plotted separately on a boxplot, shown in Figure 3.9. In this figure, no outliers below the minimum of the boxplot were observed. The farthest point above the maximum of the boxplot was for the title "Understanding the Global Energy Crisis", which occurred in August 2014 and represents 569 downloads. Some other outliers observed were "China's iGeneration" in March 2017, with 352 downloads; "On Global Citizenship" in October 2016, with 460 downloads, and October 2015, with 341 downloads; and "Constructing Muslims in France" in November 2015, with 308 downloads.

To compare the access of each title to the access of English-language titles on OAPEN, a distribution graph was also plotted. The distribution of the English-language titles followed a power graph, while the data for each of the titles followed a normal distribution.

To obtain an overview of the performance of each title, benchmarking was done using subject field normalisation. First, all of the English-language titles published after 2010 were extracted. Then, each KU pilot collection title's subject was determined using the OAPEN repository metafile, and titles uploaded before July 2016 were extracted for each subject. The average monthly downloads for these subjects were then computed. One of the issues encountered was that some subjects were too specific or too general to compare: for some subjects, there was an insufficient number of titles, and it was also not feasible to compare titles with broad subjects such as history or law.

Starting from the first upload time, which was March 2014, the number of titles was insufficient to perform a comparison. For this reason, a one-year period occurring as recently as possible had to be found. Thus, a 12-month period from the beginning of July 2016 to the end of June 2017 was selected. Another challenge was that some of the titles belong to more than one subject category. For these cases, the most specific subject containing at least ten titles was chosen, so that there would be a sufficient number of titles to perform the comparison. The titles and their subject categories are shown in Table 3.6. Figure 3.10 displays boxplots illustrating the average

monthly downloads of 11 titles alongside the monthly downloads of titles with the same subject.

Table 3.6: Subject categories of 11 titles.

Title	Subject Category
1. Aging Gracefully	Literature & literary studies
2. Biological Relatives	Politics & government
3. China's iGeneration	Media studies
4. Constructing Muslims in France	Political science & theory
5. Electronic Iran	Media studies
6. Networks and Institutions in Europe's Emerging Markets	Politics & government
7. On Global Citizenship	Political science & theory
8. Partisan Gerrymandering and the Construction of American Democracy	Political science & theory
9. The Ethics of Armed Conflict	Political science & theory
10. The Myth of Piers Plowman	Literature & literary studies
11. Understanding the Global Energy Crisis	Political science & theory

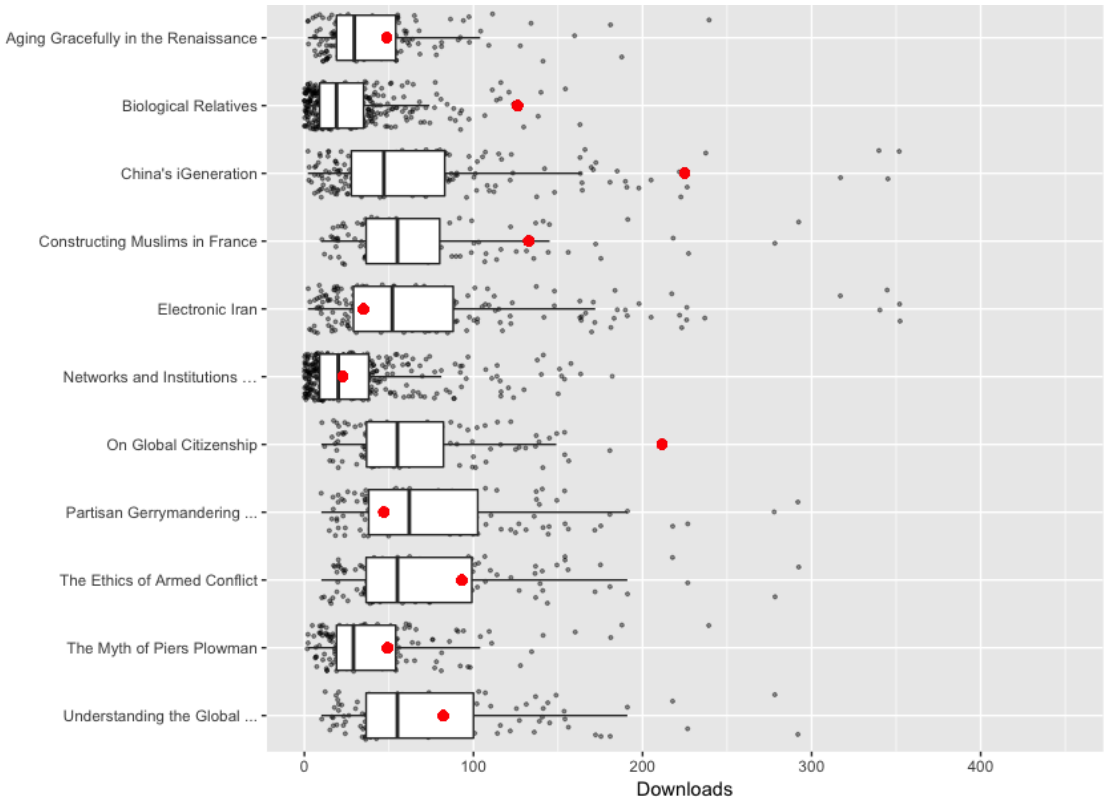

Figure 3.10: Average monthly downloads of each title (represented by a red dot) alongside the monthly downloads of titles with the same subject.

In Figure 3.10 the average monthly downloads of the first seven titles (represented by a red dot) are within the interquartile range (IQR) of the average monthly downloads of books with the same subject. Four titles were downloaded more than their subject IQR: “Biological Relatives”, “China’s iGeneration”, “Constructing Muslims in France”, and “On Global Citizenship”.

Three titles have monthly downloads that fall outside of their subjects’ 95% confidence intervals. According to the webometrics data presented in Table 3.5, two of these three titles were also included in the top three most visible titles on the web. The averages reported in Figure 3.10 were only for the period from the beginning of July 2016 to the end of June 2017. To check if there was a correlation between the domain presence numbers and the download numbers, each title’s average monthly downloads from the date of upload until the end of June 2017 was calculated (shown in Table 3.7). These figures were then plotted against the webometrics findings, as shown in Figure 3.11. A Spearman’s rank correlation coefficient of 0.52 was obtained ($p=0.005$), indicating a positive correlation between a title’s average monthly downloads from OAPEN and its domain presence numbers.

Table 3.7: Domain presence and average monthly downloads of each KU collection title from date of upload to the end of June 2017.

Title	Domain Presence	OAPEN Average Monthly Access
Biological Relatives – IVF, Stem Cells and the Future of Kinship	150	130
Constructing Muslims in France	108	126.7
Law, Liberty, and The Pursuit Of Terrorism	107	75.3
China's iGeneration	106	128.1
Ever Faithful	105	83.6
My Voice Is My Weapon	103	114.5
The Ethics of Armed Conflict	95	85.1
The Myth of Piers Plowman	93	73.4
Composing the Party Line	90	23.6
Human Rights and Democracy	90	95.8
Oaths and Swearing in Ancient Greece	88	97.2
The Emergence of Irish Gothic Fiction	84	105.5
Fighting for a Living	83	84.2
Beastly Journeys	79	51.2
On Global Citizenship	79	138.5
Partisan Gerrymandering and the Construction of American Democracy	78	37.9

Title	Domain Presence	OAPEN Average Monthly Access
Passionate Amateurs	76	69.4
Networks and Institutions in Europe's Emerging Markets	72	32.2
Understanding the Global Energy Crisis	70	128
Aging Gracefully in the Renaissance	67	53.8
Making and Unmaking in Early Modern English Drama	62	43.7
Rhetorics of Belonging	62	101.4
Thinking and Killing	60	58.2
The World Jewish Congress During The Holocaust	45	19.8
In Search of the Amazon	39	85.4
Electronic Iran	35	63.2
Governing Failure – Provisional Expertise and the Transformation of Global Development Finance	34	49.4
Verse and Transmutation	28	43.3

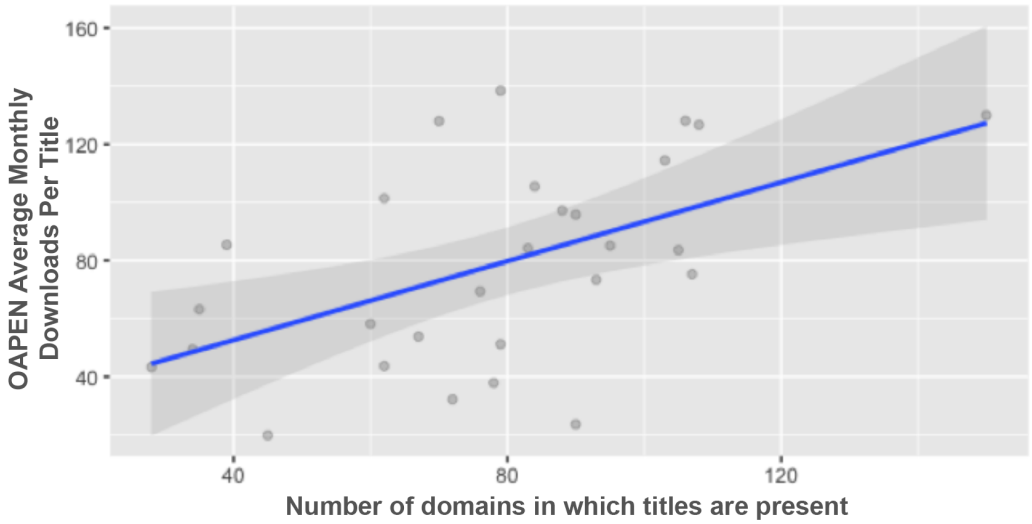

Figure 3.11: Correlation between number of domains in which the KU titles are present and the average monthly downloads of these titles on OAPEN.

In the following sections, one of these four highly visible and frequently used titles will be investigated as a study case to understand why the title was downloaded and was more visible than other titles. The investigation will also consider whether their visibility on the web directs traffic to the repositories.

Country-based downloads

Figure 3.12 shows the 40 countries with the most downloads of KU pilot collection titles, based on the OAPEN country-based access statistics.

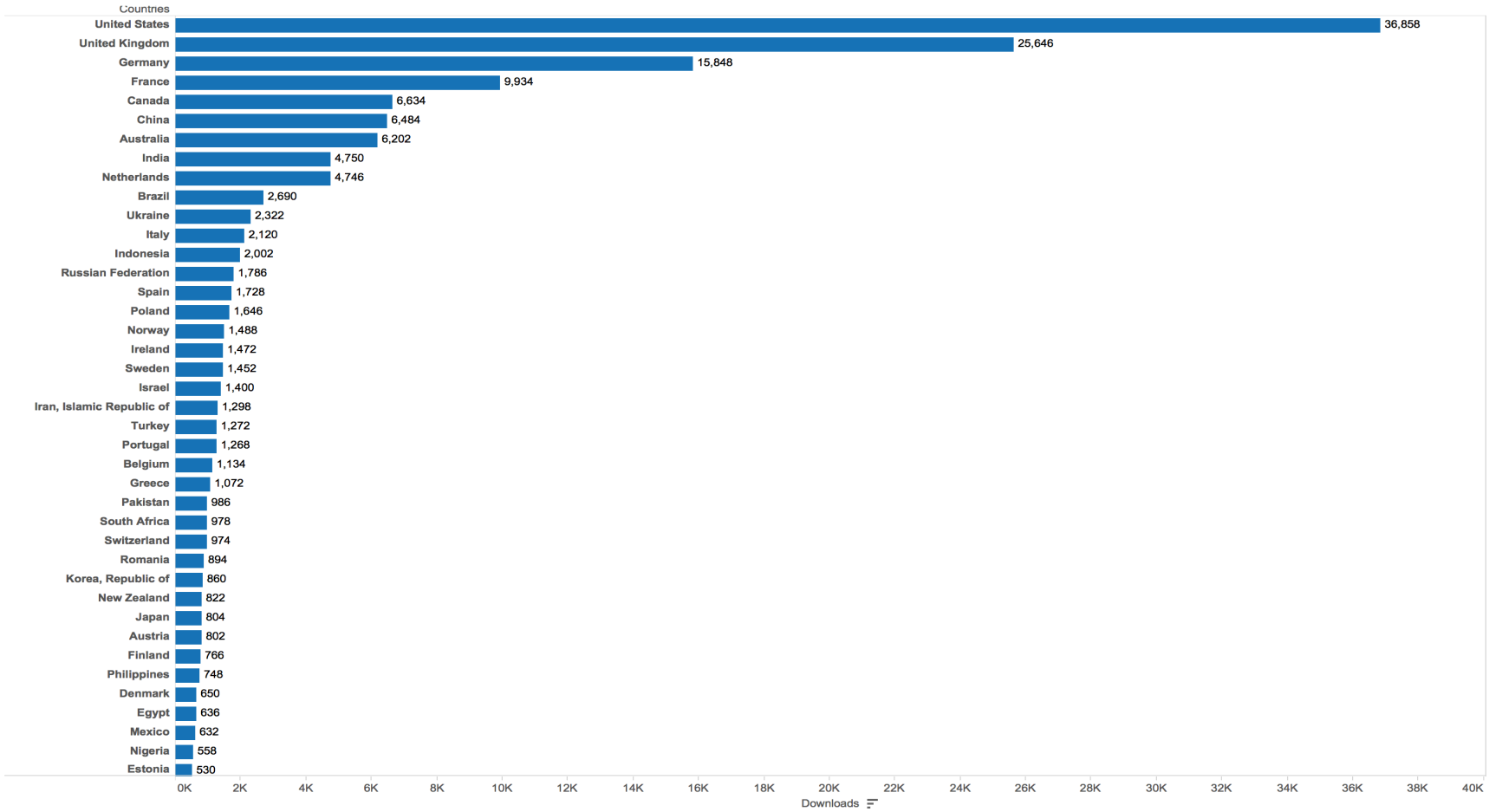

Figure 3.12: Forty countries with the most downloads of KU pilot collection titles from the OAPEN repository. This figure is similar to Figure 3.2, which shows the top-level domains with the highest number of URLs in which the 28 titles were present.

The top six countries are those that have the most universities in the top 300 globally, according to the Academic Ranking of World Universities (ARWU). They are also amongst the ten largest economies in the world. The list seems to be dominated by English-speaking countries. If the countries with the most downloads of specific titles are examined, it is observed that titles were downloaded more from the regions on which they focus. In Table 3.8, Iran is second in the list of top-downloading countries for “Electronic Iran”. Overall, Iran was ranked 32nd. Brazil (ranked 17th overall) ranked third in downloads for “In Search of the Amazon”.

Table 3.8: OAPEN country-based access of titles dealing with specific regions.

Countries

Titles	United States	United Kingdom	Germany	France	Canada	China	Australia	India	Netherlan..	Brazil	Ukraine	Italy	Indonesia	Russian Federation	Spain	Poland	Norway	Ireland	Sweden	Israel	Iran, Islamic Republic ..	Turkey
Electronic Iran - The Cultural Politics of an Online Evolution	481 (1) (Overall 1)	226 (4) (Overall 2)	227 (3) (Overall 3)	69 (7) (Overall 4)	104 (6) (Overall 5)	169 (5) (Overall 6)	38 (12) (Overall 7)	59 (10) (Overall 8)	68 (8) (Overall 9)	18 (21) (Overall 10)	67 (9) (Overall 11)	23 (17) (Overall 12)	28 (14) (Overall 13)	41 (11) (Overall 14)	10 (29) (Overall 15)	13 (24) (Overall 16)	14 (22) (Overall 17)	12 (26) (Overall 18)	24 (16) (Overall 19)	28 (14) (Overall 20)	255 (2) (Overall 21)	23 (17) (Overall 22)
In Search of the Amazon - Brazil, the United States and the Nature of a Region	1,043 (1) (Overall 1)	168 (4) (Overall 2)	604 (2) (Overall 3)	113 (5) (Overall 4)	92 (7) (Overall 5)	107 (6) (Overall 6)	42 (12) (Overall 7)	37 (14) (Overall 8)	46 (11) (Overall 9)	287 (3) (Overall 10)	60 (9) (Overall 11)	23 (19) (Overall 12)	8 (34) (Overall 13)	34 (16) (Overall 14)	22 (20) (Overall 15)	18 (21) (Overall 16)	51 (10) (Overall 17)	2 (66) (Overall 18)	17 (23) (Overall 19)	10 (30) (Overall 20)	() (Overall 21)	12 (27) (Overall 22)
Oaths and Swearing in Ancient Greece	1,034 (1) (Overall 1)	505 (2) (Overall 2)	169 (4) (Overall 3)	69 (8) (Overall 4)	126 (7) (Overall 5)	46 (11) (Overall 6)	201 (3) (Overall 7)	9 (37) (Overall 8)	50 (9) (Overall 9)	38 (13) (Overall 10)	13 (33) (Overall 11)	131 (6) (Overall 12)	16 (25) (Overall 13)	33 (15) (Overall 14)	50 (9) (Overall 15)	43 (12) (Overall 16)	38 (13) (Overall 17)	15 (28) (Overall 18)	20 (21) (Overall 19)	30 (16) (Overall 20)	9 (37) (Overall 21)	28 (17) (Overall 22)
The Emergence of Irish Gothic Fiction - Histories, Origins, Theories	956 (1) (Overall 1)	657 (2) (Overall 2)	264 (4) (Overall 3)	139 (6) (Overall 4)	115 (9) (Overall 5)	139 (6) (Overall 6)	159 (5) (Overall 7)	68 (14) (Overall 8)	53 (15) (Overall 9)	46 (18) (Overall 10)	130 (8) (Overall 11)	95 (10) (Overall 12)	34 (19) (Overall 13)	33 (20) (Overall 14)	88 (11) (Overall 15)	69 (13) (Overall 16)	86 (12) (Overall 17)	403 (3) (Overall 18)	26 (22) (Overall 19)	7 (47) (Overall 20)	17 (30) (Overall 21)	18 (29) (Overall 22)
Total KU Downloads (Pilot)	18,429 (1) (Overall 1)	12,823 (2) (Overall 2)	7,924 (3) (Overall 3)	4,967 (4) (Overall 4)	3,317 (5) (Overall 5)	3,242 (6) (Overall 6)	3,101 (7) (Overall 7)	2,375 (8) (Overall 8)	2,373 (9) (Overall 9)	1,345 (10) (Overall 10)	1,161 (11) (Overall 11)	1,060 (12) (Overall 12)	1,001 (13) (Overall 13)	893 (14) (Overall 14)	864 (15) (Overall 15)	823 (16) (Overall 16)	744 (17) (Overall 17)	736 (18) (Overall 18)	726 (19) (Overall 19)	700 (20) (Overall 20)	649 (21) (Overall 21)	636 (22) (Overall 22)

The coefficient of the Spearman's rank correlation between downloads per country from OAPEN and the number of web resources for each TLD, in which the 28 KU titles were present, was 0.66 ($p < 0.0001$). However, the coefficient of the Pearson correlation was found to be 0.96 ($p < 0.0001$), which suggests that although the rank is moderately positive, the linear correlation is much larger. Because the linear correlation is driven by the three large number. These results indicate that there is a positive correlation between a country's downloads of a title and the number of web pages hosted in that country that mention the title. The relationship between web presence per country and OAPEN downloads per country is shown in Figure 3.13.

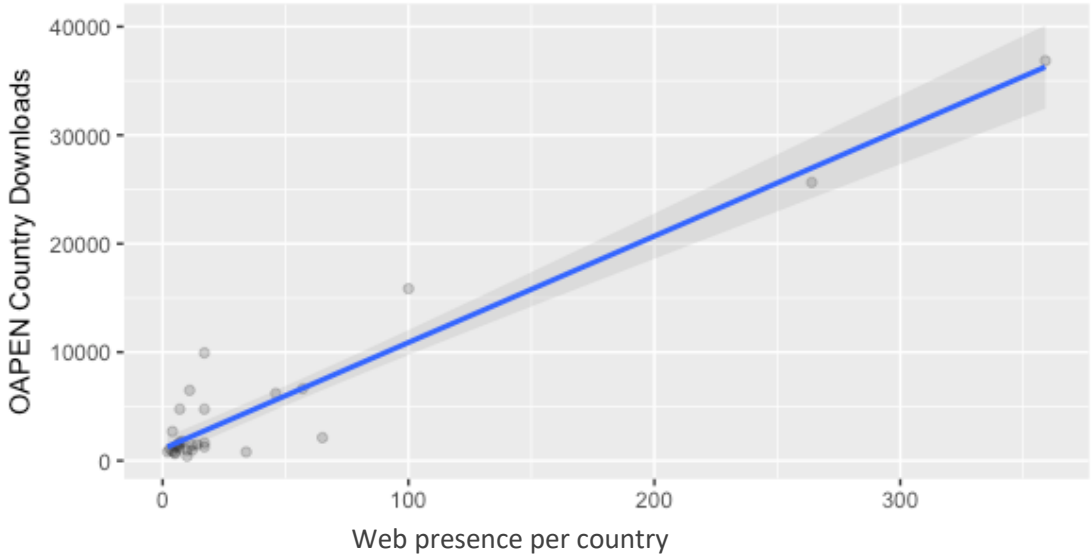

Figure 3.13: Web presence per country with respect to downloads per country from OAPEN for each title.

However, after removing the three large driving numbers and reanalysing the data, the Pearson correlation coefficient was reduced to 0.62, and the Spearman's rank correlation coefficient remained almost the same, at 0.61, both with $p < 0.0001$. This suggests that the Spearman rank correlation is the more relevant analysis in this case, where a few outliers can exert sufficient influence to change the correlation coefficient.

3.5.5.2 COUNTER-compliant IP address access

3.5.5.2.1 Institutional Access

To obtain institutional access figures, the pledging libraries for the KU collections were asked for their institutional IP address blocks. Subsequently, the IP address accesses of KU collection titles in the OAPEN repository were collected for the

period from the beginning of March 2014 to the end of March 2017. Subsequently the IP address blocks of 362 institutions were extracted from these IP address-based download reports.

Figure 3.14 shows that institutional downloads of the KU pilot collections via KU-pledging libraries represent approximately 15% of the total KU collection downloads. This suggests that pledging libraries are making significant use of the knowledge contained in specialist scholarly books. The average number of titles downloaded per pledging institution is 33.4 over three years of access. This does not include the university/library members that download titles outside of the university's campus or premises (Knowledge Unlatched, 2017).

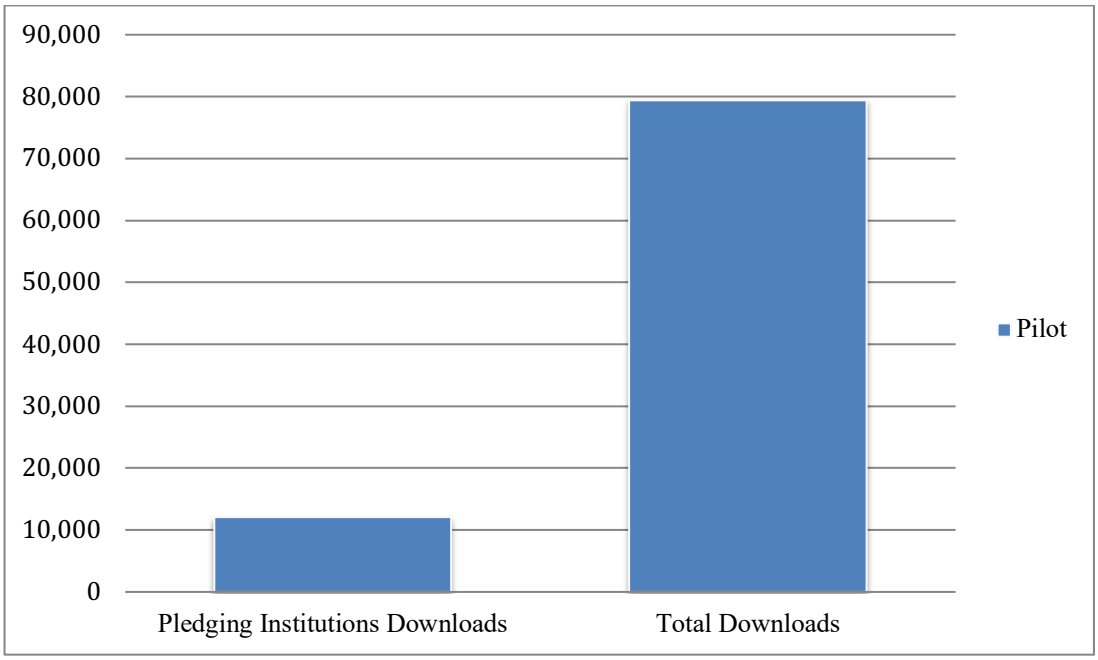

Figure 3.14: Pledging libraries' downloads vs. total downloads of the KU Pilot collections.

3.5.5.2.2 Geolocating downloads

Using IP address-based COUNTER-compliant reports from OAPEN for the period of March 2014 to June 2017, repository access for the KU pilot collection titles was geolocated, as shown in Figure 3.15. It can be seen that these titles are mainly accessed from Europe, the eastern and western parts of the United States, and Australia, China, and India.

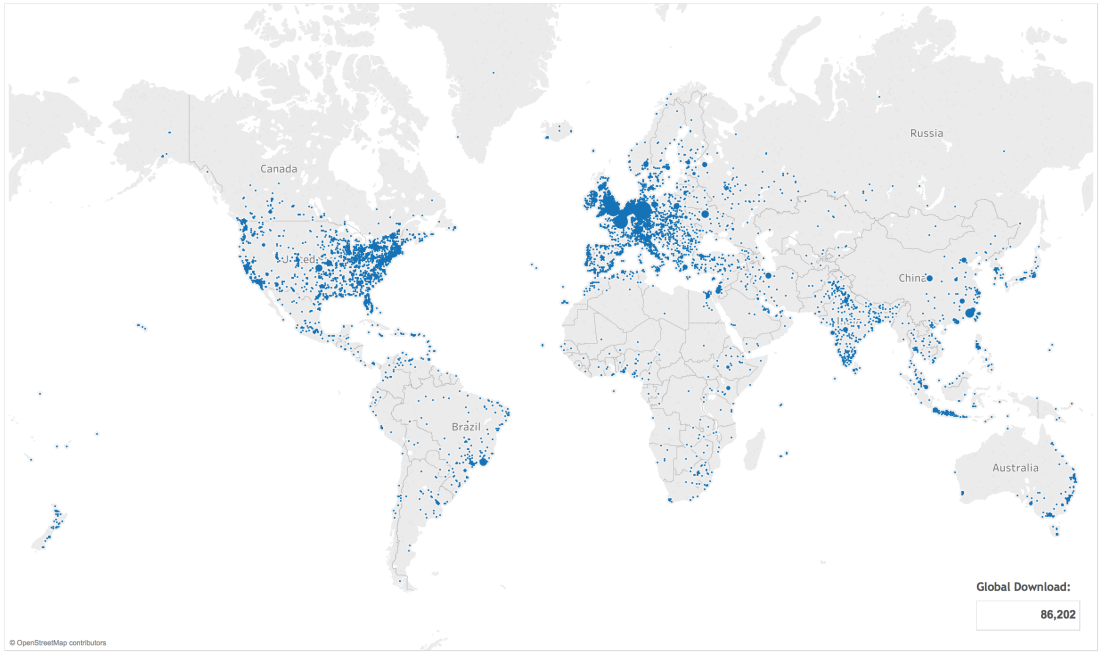

Figure 3.15: Access geolocations for the KU pilot collection on OAPEN.

3.5.5.2.3 State-based access

By reverse geolocating the latitudes and longitudes obtained in the previous section, access was investigated according to administration area level 1, which corresponds to states in the United States and Australia and to the four countries in the United Kingdom. Figure 3.16 shows the state-based access distribution in the United States. Most accesses came from California and New York. This access map looks similar to a population distribution graph of the United States, with some differences, as the access from New York is higher than that from Texas. A Spearman's rank correlation coefficient of 0.83 ($p < 0.0001$) was obtained, indicating a high positive correlation between downloads from OAPEN and US state populations.

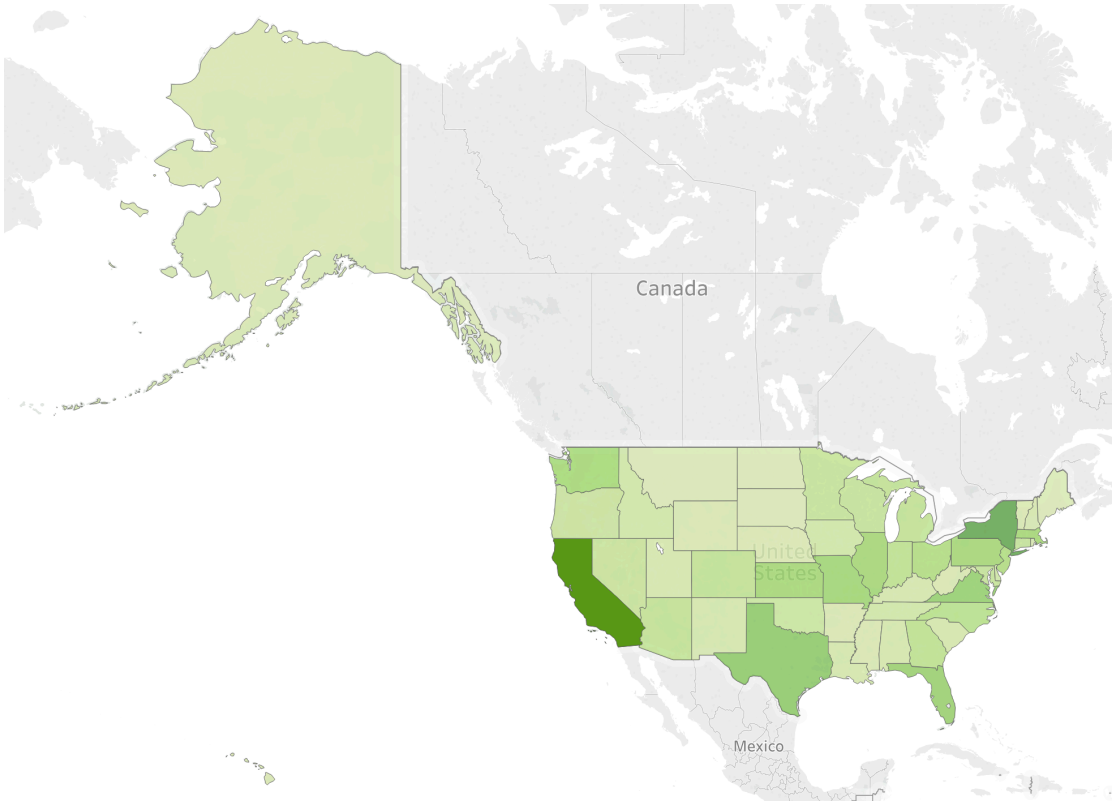

Figure 3.16: State-based access distribution of KU pilot collection titles in the United States.

3.5.5.3 Web analytics: Page view metrics

In addition to identifying readers' geographical origins, web page analytics data were needed to identify the sites on which they were discovering these titles. For this purpose, page view statistics for the presentation web pages of each title were examined, where each page has a direct link to the PDF file of the relevant book. To examine web access statistics, web page access first had to be checked to determine if it was high enough to reflect downloads. Since OAPEN uses Google Analytics to track web access, the 'RGA' r package was used to download website statistics.

After website statistics were downloaded from OAPEN, unique page view numbers were compared to download figures. For the period between March 2014 and June 2017, the total number of downloads of the 28 titles was 86,202, with 39,739 unique page views for these 28 titles. This suggests that at least half of the readers downloaded the titles without visiting the OAPEN web pages. Figure 3.17 shows both downloads and unique page views by month.

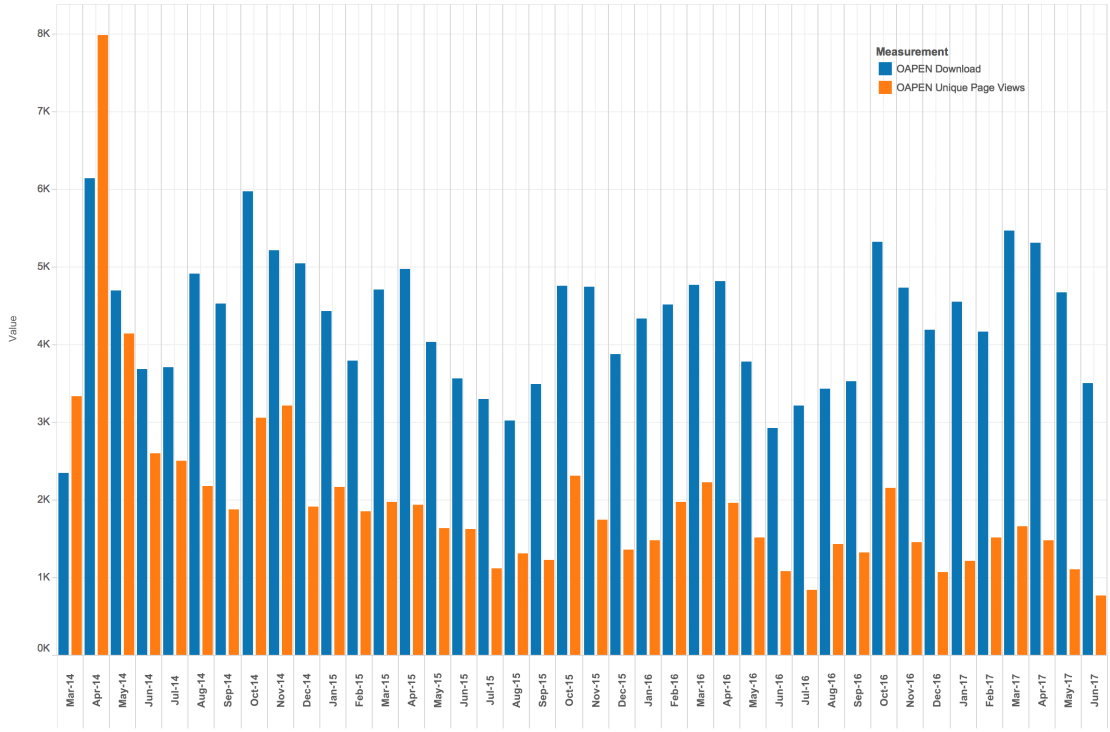

Figure 3.17: OAPEN unique page views vs. OAPEN downloads for KU pilot collection.

Figure 3.17 shows that unique page views were higher than downloads until May 2014, after which the downloads per month were considerably higher than the monthly page view numbers. This suggests that users were downloading the titles without visiting the presentation pages on the repository. A moderate positive Pearson correlation ($r=0.46$ with $p < 0.0001$) was found between downloads from OAPEN and unique page views. Since the number of unique page views represents at least one third of the number of downloads and there is a positive correlation between the two, data from Google Analytics can be used to give an indication of where readers were coming from and which channels were more effective in the dissemination of the titles.

3.5.5.4 Web analytics: Session (visit) metrics

According to Google Analytics, there were a total of 28,205 referral visits coming through sources other than search engines for all 28 KU collection titles at OAPEN. The top traffic source for all 28 books is direct links (7,387 sessions, 26.2%). Direct links involve users typing the URL into a browser or clicking on a bookmark. Unfortunately, in this case, Google Analytics cannot determine where the user comes from, because there is no information in the HTTP referrer header. The

second-highest referrer is the knowledgeunlatched.org site (with 3,938 sessions, 14%), which lists all the titles in the collection and directs readers to OAPEN for the PDF downloads. The third and fourth top sources are e-booksdirectory.com (1,900 sessions, 6.74%), which provides a daily list of free downloadable e-books, and onlinebooks.library.upenn.edu (1,081 sessions, 3.8%), which lists over 2 million free books on the web. Since these are free listing sites, which, unlike library catalogues, can be freely accessed by anyone, the titles they referred to were checked. The e-booksdirectory site, which lists 10,383 books as of January 2018, referred to only six titles from the KU pilot collection, namely "Biological Relatives", "Governing Failure", "Human Rights and Democracy", "The Ethics of Armed Conflict", "Thinking and Killing", and "Understanding the Global Energy Crisis". This list suggests that e-booksdirectory features titles that are interesting to a more general audience outside of academia, which is plausible, because people at universities mostly use library catalogue search engines, Google Books searches, and general search engines such as Google or Bing rather than free listing sites (Springer, 2010). In addition, even in general search engines, when the title of a KU monograph is entered, the resulting page displays the publisher's site and the OAPEN repository site above e-booksdirectory.com and onlinebooks.library.upenn.edu. This suggests that people coming from these sites are in fact followers of these free listing sites and likely learned about these titles from these sites and not elsewhere.

Google Analytics shows referrals by host name. Therefore, different host names belonging to the same company or organisation can be identified. For example, users come from different country hosts of Google Scholar, such as scholar.google.com, scholar.google.com.au, and scholar.google.co.uk. These hosts were grouped under one domain name (i.e., scholar.google). By filtering domain names, 1,242 sessions (4.4%) that came from ProQuest's Serials Solutions system were found. This system provides libraries with e-resource access and management services. Another ProQuest company, ExLibris, which provides library systems, was the source of 1,156 sessions (4.1%). EBSCOHost, another library system, was the source of 327 sessions (1.2%). Google domains were the source of 1,398 sesions (5.0%), 603 (2.1%) of which came from Google Scholar. Most of the hosts from universities were identified as library catalogue services or EZproxy systems, which are web

proxy servers that library systems use. These web proxy servers provide access from outside the library's computer network to restricted-access websites that authenticate users according to their IP address. Therefore, session numbers shows that library cataloguing and search systems still play a more important role in the discovery of monographs than Google Scholar does.

3.5.5.5 Web analytics: Social network referral metrics

In order to determine how effective social networks are in directing users to monograph pages, social network sources were examined in Google Analytics. The vast majority of social network sources of the 31,138 unique page views were in the "not set" category. Google explains that "not set" denotes any direct visit or referral visit where the link does not have a keyword, ad content, or any other suffix in the URL with any campaign information associated with the visit (Google, 2009). In other words, these "not set" referrals are not defined in the Google Analytics default channel grouping.

Of the unique page views that came from social network sources, 477 came from Facebook and 324 came from Twitter. Table 3.9 shows that the total traffic directed from social networks makes up only 2.71% of the total number of unique page views.

Table 3.9: Social network sources for page views of the 28 KU titles' web pages on OAPEN.

Social Network	Unique Page Views	Share
(not set)	30,293	97.29%
Facebook	477	1.53%
Twitter	324	1.04%
Tumblr	27	0.09%
Blogger	14	0.04%
StumbleUpon	1	0.00%
Digg	1	0.00%
Academia	1	0.00%

3.5.5.6 HathiTrust repository access

Another platform that hosts KU pilot collection titles is HathiTrust. HathiTrust, founded in 2008, is a partnership of major research institutions and libraries that hosts digital content from research libraries, including content digitised via the

Google Books project and the Internet Archive digitisation initiative. As of January 1st, 2017, they had more than 14 million volumes in their collection, 5.7 million of which were in the public domain (Zaytsev, 2017). HathiTrust began hosting KU pilot collection titles in March 2014.

Figure 3.18 shows HathiTrust's monthly unique page views alongside OAPEN's monthly downloads. Compared to the OAPEN downloads, HathiTrust page views are few in number, and there is no obvious correlation between the two sets of figures. In order to understand how people were discovering these titles, the source of the HathiTrust traffic was examined.

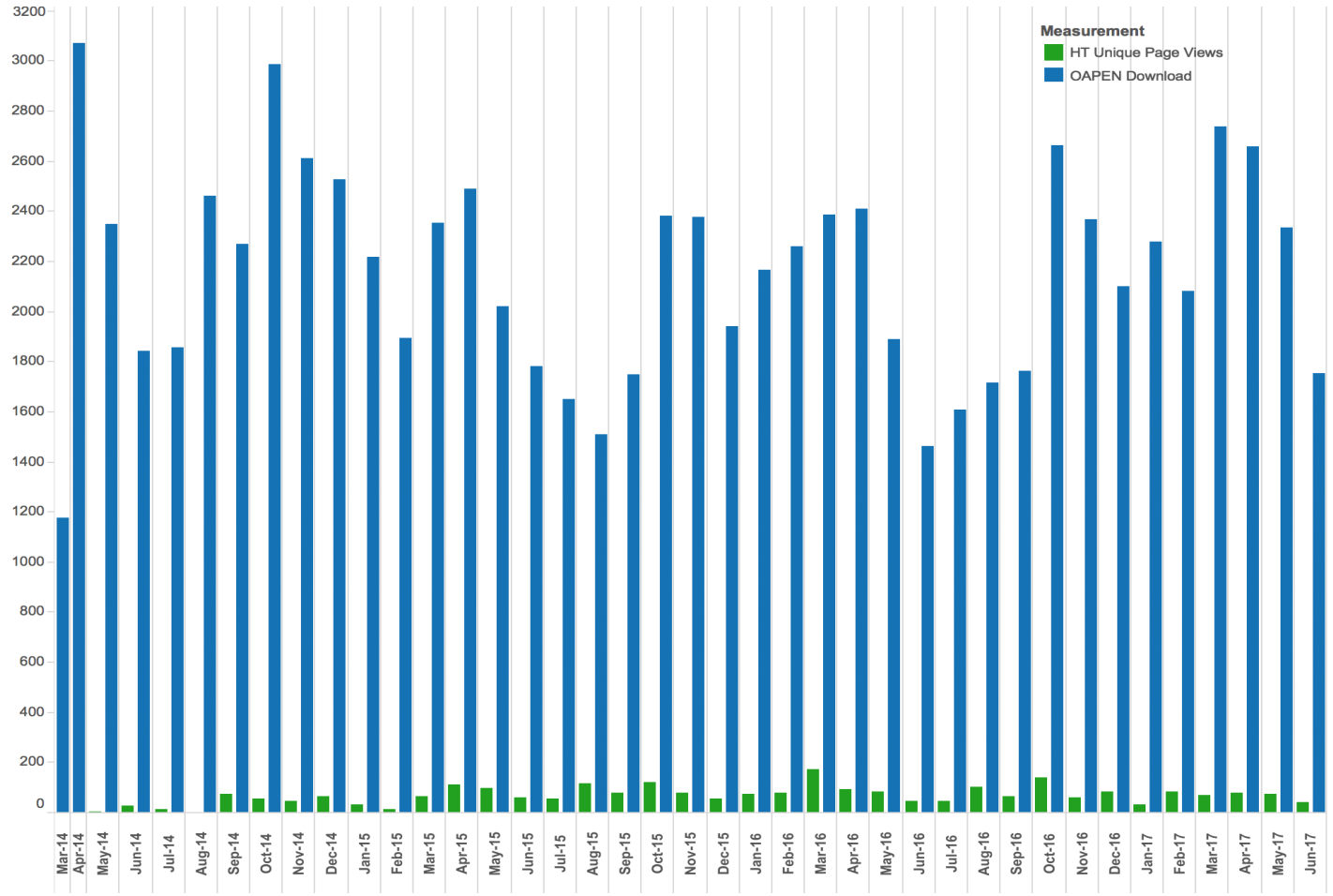

Figure 3.18: HathiTrust unique page views vs. downloads from the OAPEN for all KU titles.

Only 12 referrers could be identified from Google Analytics for the entire KU pilot collection on HathiTrust. HathiTrust's lower page view numbers were also due to the lack of library management and cataloguing systems that link to HathiTrust. All of these systems link to the OAPEN repository instead.

HathiTrust provides PDF downloads in addition to online web views of the books. No comparison was made between PDF downloads and web views, as there were very few downloads. One of the reasons for the lack of downloads from HathiTrust is that when downloading the PDF file of the book, the file is built by converting several pages to PDF format and binding these together. Each page has a HathiTrust watermark, which increases the file size. It may be that users prefer to download the original, unchanged PDF format from OAPEN or other repositories.

In conclusion, titles are accessed more on the OAPEN repository than the HathiTrust repository for various reasons. First, the titles on OAPEN are registered in the DOAB, which makes them more discoverable through library catalogues. Moreover, the web pages on which these titles are present, although they were few, provided links to the OAPEN repository instead of the HathiTrust repository. Therefore, when queried on web search engines, titles on the OAPEN repository appeared at the top of the results page, which led to them being accessed more.

3.5.6 Discussion

The most common types of website on which the 28 KU pilot collection monographs were present were university, scam, and bookseller sites. Scam sites use OA books as bait to attract users and collect their email addresses and credit card details. Although these books are OA, there is no way for users to know that they are freely accessible on the internet. For this reason, they are still seen as a commodity and can still be used as bait.

Mainstream media sites such as *Science* magazine and *The Washington Post* have a larger audience than university and bookseller's sites. These mainstream websites, along with Wikipedia, news sites such as that of the BBC, and radio programs are also important in the dissemination of titles across society beyond academia.

A moderate positive correlation (r_S=0.52 with p=0.005) was found between the number of domains in which each title was present and the average monthly

downloads for each of these titles. If websites provided more links to the PDFs of the titles, this coefficient might increase. It is also possible that the owners of these pages did not know that the titles were OA. This may also be the case for readers outside of academia. If they do not search for the OA version of the title, there is no way for them to access the PDF version of it. One approach to eliminate this problem would be to discuss this problem with publishers, so that they do place a link to the PDF on their pages to redirect users to the OA versions of the books. Sites like Amazon.com and other vendors also do not direct users to the free version of the book, which constitutes another obstacle to readers accessing the OA versions of these titles.

A positive Spearman's rank correlation coefficient (r_S=0.66, $p < 0.0001$) was found between the number of sites based in a particular country in which these title names were present and the number of downloads from that country. Germany was in the top three of both web presence per country and downloads per country, but the number of scam sites on which these books were present was much higher in this country than in others. Unglue.it also inflated the web presence for Italy. However, it was the country with the 12th-most downloads from OAPEN.

This study did not collect data that would allow an analysis of book access from a specific geographic location to be linked to the book's subject as a result of the title's OA attribute. In future, a study comparing the access of a larger number of paywalled and OA titles from geographic locations mentioned in the books' subject metadata would help to reveal whether OA encourages greater use from these potential target audiences.

In this study, it was possible to investigate the sources of some unusual book downloads by identifying title-specific monthly downloads and understanding their causes based on page views and web referrers. By working closely with authors, publishers will be able to make sense of how authors' actions, including their appearance on TV and radio programs, their articles and posts on social media affect downloads. This would also make it possible for publishers to see the effect of their efforts to disseminate their books.

3.6 Case study: "Constructing Muslims in France"

The title "Constructing Muslims in France" by Jennifer Fredette was chosen as a case study, because, as shown in the analyses above, it was the second most visible title, with a presence in 108 domains, and it had the second-highest average monthly download rate. This title covers the diversity and complex identity politics of Muslims in France and contrasts it to framings of Muslims as failed and incomplete French citizens in French media and elite public discourse. It is probable that this monograph attracted attention because it deals with subjects of current interest. For example, after the terrorist attacks perpetrated by ISIS in France, people and news sites became more interested in the issues and challenges that France is facing. These attributes of this monograph make it a good candidate for a case study, in which most of the approaches discussed in this chapter can be applied to it to obtain a detailed picture of its visibility and access.

Regarding its visibility, "Constructing Muslims in France" was present on 162 web resources, and the URL analysis and content analysis revealed that the majority of resources in which this title name was present were from academic domains (25%), more than half of which were from the United States (13.6%). The second most common resources were scam sites (16%), followed by repositories and publishing platforms. One of these repositories, BiblioVault, which serves more than 90 scholarly publishers in the US and Europe, does not provide a link to the PDF file, but only to the publisher's site, where only paid versions of the title can be found. Some forum sites on islamophobia, including islamophobiawatch.co.uk and the Council for European Studies, also shared this title on their pages.

The title "Constructing Muslims in France" is also present on three mainstream news sites, namely *The Washington Post*, *Huffington Post,* and *The Telegraph India*. *The Telegraph India*, which is the fifth most widely read English newspaper in India, discusses the book without providing a link to it. The relevant article was published just after the Charlie Hebdo shooting, which occurred in France on 7 January 2015. The *Huffington Post* article is also on the Charlie Hebdo attacks. In *The Washington Post*, two articles written by the book's author, Jennifer Fredette, were published on July 29, 2014 and October 2, 2014. Both include links to the book's page on Fredette's own website, where she provides links to the PDF file of

the title hosted on OAPEN and to the title's page on the publisher's site. The article published on July 29 also gives a link to the PDF file of the title hosted on OAPEN. All of the mainstream news sites that discuss the book were in English, which is the language in which "Constructing Muslims" is written.

Events—in this case, terrorist attacks—can also trigger monograph downloads. For this title, peak downloads (302 downloads) occurred just after the Paris terrorist attacks on 13 – 14 November 2015. This was the single deadliest terrorist attack in French history, causing 130 deaths and injuries to 368 people. Figure 3.19 shows downloads and unique page views on the OAPEN platform for "Constructing Muslims in France".

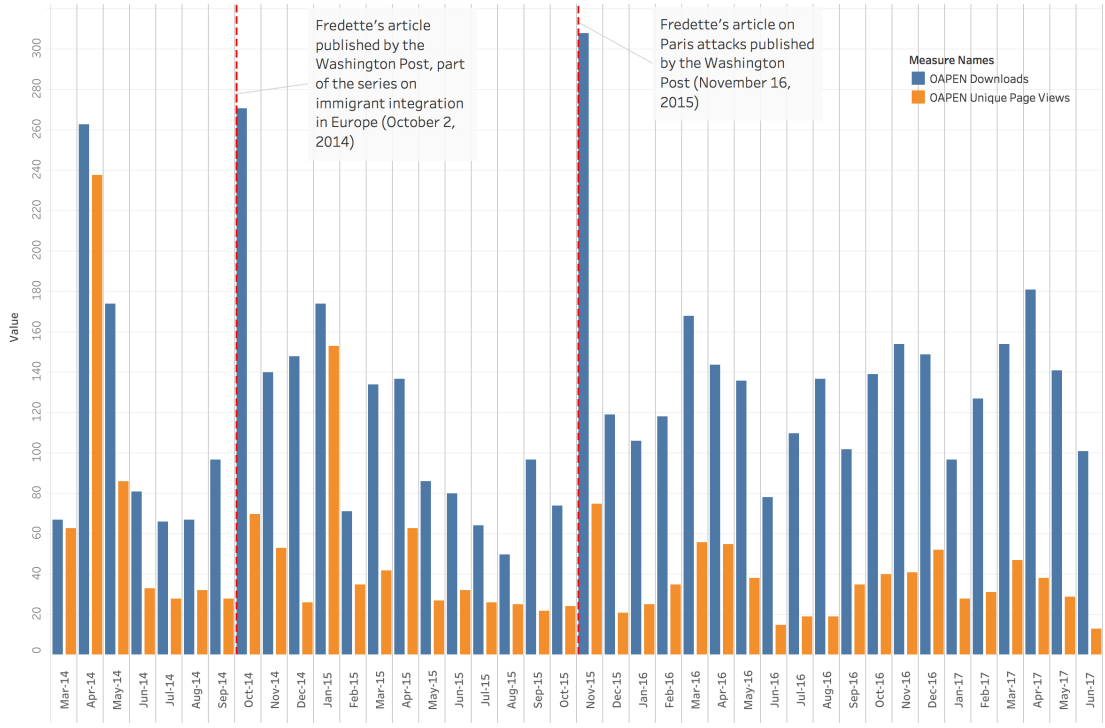

Figure 3.19: Peak downloads on the OAPEN platform for "Constructing Muslims in France" occurs in November 2015 after Fredette's article was published on the Washington Post.

Figure 3.19 shows that most of the downloads of the book occurred directly, without the downloader visiting the web page. Although the web page access figures are low, Google Analytics data showed that the top 10 referring sites for the month of November 2015, besides library catalogue sites, included Jennifer Fredette's blog (http://jenniferfredette.com); Facebook; and *The Washington Post* site, where Fredette had published an article on the Paris terror attacks in which she cited her book (Fredette, 2015). The country-based download counts for the month of November

revealed that most of the 302 downloads in this month were from the United States (195), followed by the United Kingdom (47) and Canada (31). Table 3.10 shows the ten countries with the most OAPEN downloads of "Constructing Muslims" in November 2015 compared to these countries' download counts for the period of March 2014 – July 2017. The United States' share in November 2015—51.6%—is twice that of its share for March 2014 – July 2017. The IP address-based download reports make it possible to geolocate downloads in the United States for November 2015 (Figure 3.20).

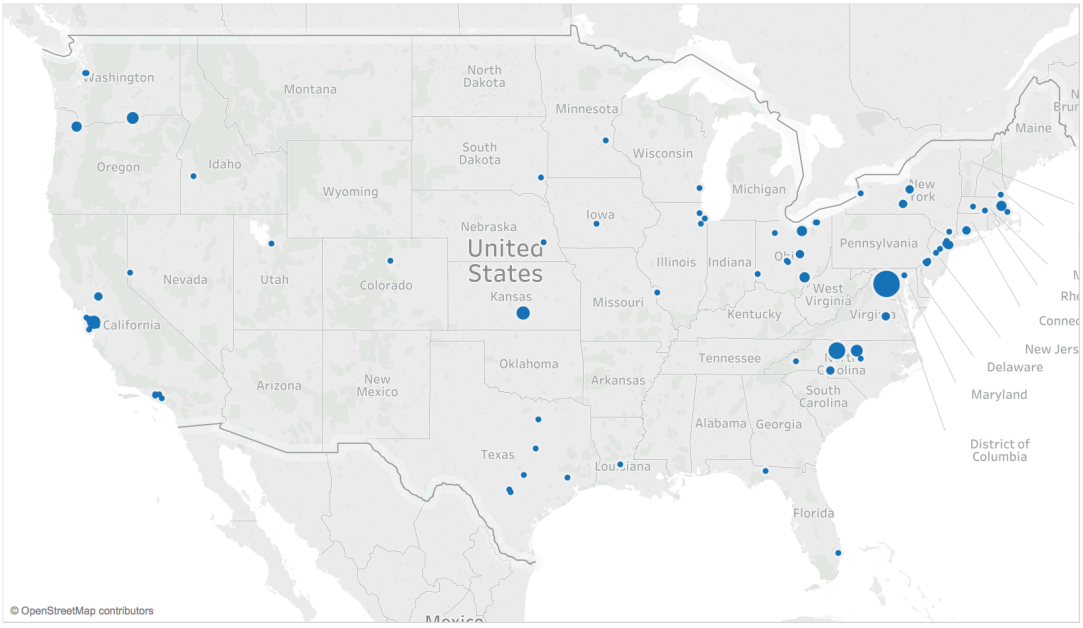

Figure 3.20: The geolocations of US downloads of "Constructing Muslims in France" in November 2015.

Figure 3.20 shows that November 2015 downloads in the United States were not concentrated in a few locations, which suggests that the high download figures were not caused by the book being prescribed for a university course or another local event. Thus, it is possible to conclude that the title's visibility in *The Washington Post* may be a cause of this spike in downloads from the United States (shown in Table 3.10).

Therefore, it would not be wrong to say that mainstream media sites have an important effect in increasing downloads of a title.

Table 3.10: Ten countries with the most OAPEN downloads of "Constructing Muslims in France" in November 2015 and the period of March 2014–June 2017.

	November 2015			March 2014–June 2017		
Countries	**Downloads**	**Share**	**Rank**	**Downloads**	**Share**	**Rank**
United States	159	51.62%	1	1195	25.50%	1
United Kingdom	23	7.47%	2	647	13.80%	2
Canada	22	7.14%	3	186	3.97%	6
France	14	4.55%	4	397	8.47%	3
Ukraine	12	3.90%	5	77	1.64%	10
Australia	9	2.92%	6	147	3.14%	7
Germany	8	2.60%	7	271	5.78%	4
Netherlands	7	2.27%	8	97	2.07%	9
Japan	5	1.62%	9	45	0.96%	18
Czech Republic	4	1.30%	10	22	0.47%	30

The second-highest number of OAPEN downloads of "Constructing Muslims" occurred in October 2014, when Fredette's article, which contained a link to her site, appeared in *The Washington Post*. It can be seen from the top ten referring sites shown in Figure 3.21 that the site that drove the most traffic, following the Knowledge Unlatched and serialsolutions.com domains, is Jennifer Fredette's blog, with 103 unique page views.

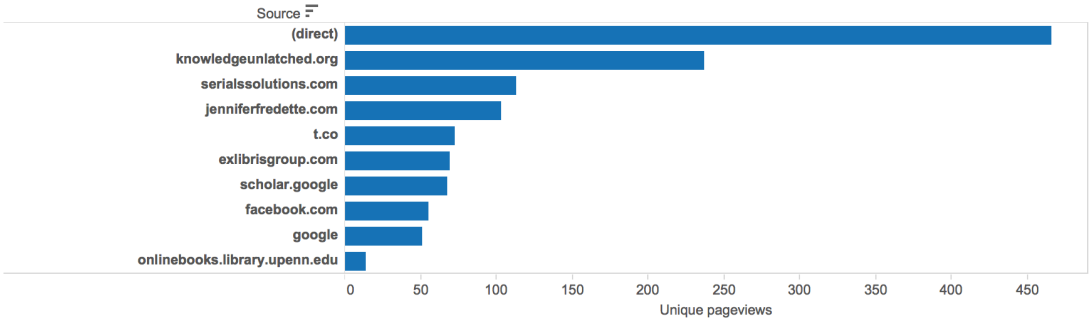

Figure 3.21: Top ten referring sources for the page of "Constructing Muslims in France" on the OAPEN repository for the period March 2014–June 2017.

3.7 Conclusion

This chapter explored the discoverability, visibility and usability of the 28 KU pilot collection titles. In the dissemination of these titles, important factors are the way in which the repository shares their records with other parties and the registration of the

titles in the DOAB. The more a repository's title appears on top of a results page the more it shows the discoverability of this repository's title. Therefore, the success of a repository in sharing their records is also reflected in the results pages of library catalogues and search engines.

The visibility analysis showed that the most visible titles on the web were those dealing with subjects of current interest. On the other hand, monographs with subjects of interest to a smaller audience were less visible on the web. The 28 KU pilot collection titles were most present on pages hosted on university sites, bookseller sites, and scam pages, respectively. The majority of the 3,238 web resources on which these titles were present were from domains in English-speaking countries, and the titles occurred most frequently on .edu domains. When country-independent TLD's were removed, these titles were still most often found in domains from the United States, the United Kingdom and Germany, respectively.

The access analysis showed that the titles were downloaded less during the summer months in the Northern Hemisphere. Other than this, no specific downloading patterns were observed in the periods after each monograph was made OA. However, the access distribution analysis revealed unusual access to some titles. These were related to specific incidents, such as the November 2015 Paris terror attacks. The access distribution analysis can also be used to measure the success of book marketing events. A positive correlation was found between the countries that downloaded the KU pilot collection titles from the OAPEN repository and the countries in which the titles were most visible.

Web analytics tools are useful in showing the traffic directing users to titles. An examination of the referring sites for the KU pilot collection titles revealed that library catalogues and search systems are more effective than Google Scholar in the discovery of these monographs. Traffic occurring from social network platforms to the titles' web pages on OAPEN was limited: Facebook had a 1.5% share and Twitter a 1.0% share of the total unique page views.

In this chapter, data relating to the discoverability, visibility and access of the KU pilot collection titles was captured using standard low-cost tools and services. In the following chapter, altmetrics methods and more sophisticated tools and services are used to collect data in order to provide a deeper understanding of access to these

monographs. In this analysis, some other aspects of these titles are examined, such as social network mentions, references on Wikipedia, and citations.

4 An Analysis of Social Media and Citation Data on OA Monographs

4.1 Introduction

In the previous chapter, the discoverability of the 28 KU pilot collection titles was investigated by analysing registration in directories and indexes, repositories' record quality, and identification standards including ISBN and DOI. Monograph visibility on the web was examined using webometrics methods, and access was investigated using access data, including web page views and PDF downloads. These were analysed using standard software that is freely available. Correlation analyses showed associations between these data. Specific instances of access were investigated with the goal of understanding the causes of high access.

This chapter presents an analysis of the social media mentions and citations of these titles. First, to understand how the titles are discussed on social networks, mentions of these titles on Facebook and Twitter are considered. A range of ways to capture and interpret these data is explored. Citations from blog pages and the free encyclopaedia and general reference platform Wikipedia were collected and analysed. Review and rating data were collected from three book platforms: Google Books, which offers full texts of books and also links to various seller sites; Amazon, the largest online bookseller in the world; and Goodreads, the world's largest site for readers and book recommendations. To gather information about researchers who were reading these books, bookmark data were collected from Mendeley, a platform for managing personal reference collections and sharing research outputs.

To gain insight into how these books are read, the annotation services Hypothes.is and PaperHive were searched for the KU pilot collection titles. Subsequently, a number of citation databases, including Scopus, WoS and Google Scholar, were examined to understand how these titles are cited in academic publications. A correlation analysis was conducted on the metrics that had been collected up to this

point, including the webometrics and download statistics from OAPEN. Lastly, the meanings of these metrics are discussed, along with the challenges and issues encountered during their collection and analysis.

4.2 Datasets

A social network dataset was compiled for the period between 1 January 2014 and 1 July 2017. This database included 493 records from Twitter for all of the KU pilot collection titles and 96 records from Facebook for 20 of these titles. The UCL Press title "How the World Changed Social Media" was chosen to be the subject of a case study investigating the dissemination of these titles across the Twittersphere. Between 1 July 2016 and 1 April 2017, 181 tweets from 103 users referring to this title were captured.

The Wikipedia dataset also included 23 articles referring to 13 monographs from the KU pilot collection. All of the 28 titles' pages were collected from Google Books, Amazon and Goodreads. The Goodreads datasets also includes 49 ratings for 16 titles. The Mendeley dataset includes 20 titles, which were bookmarked by 288 readers, whose academic status is visible. The citation data includes eight indexed titles from Scopus, 28 indexed titles from Google Scholar and 27 indexed titles from WoS. The datasets are summarized in Table 4.1.

Table 4.1: Summary of collected data.

Data source	Date	Data
Twitter	1 January 2014–1 July 2017	493 records for 28 titles
Facebook	1 January 2014–1 July 2017	96 records for 20 titles
Twitter for one UCL title	1 July 2016–1 April 2017	181 tweets for one title
Wikipedia	Until March 2018	23 articles referring to 13 monographs
Goodreads	Until March 2018	49 ratings for 16 titles
Mendeley	Until March 2018	20 titles bookmarked by 288 readers
Scopus	Until March 2018	Citation data for eight titles
Google Scholar	Until March 2018	Citation data for 28 titles
WoS	Until March 2018	Citation data for 27 titles

4.3 Social media metrics and citation databases

As social media has begun to be used in scholarly communication, the visibility of scholars and scholarship has increased considerably. These social media platforms provide new ways of disseminating research outputs. These platforms allow users to connect and interact with one another by sharing and commenting on content that is created, discussed, and reused by other users. The ability to measure engagement has motivated research funders and managers to look for new metrics that are capable of demonstrating the value of research to a broader audience (Sugimoto, Work, Larivière, & Haustein, 2016). The demand for new indicators, combined with the possibilities of the social web, has led to the development of a new set of indicators called altmetrics (Priem et al., 2010), which examine users' engagement with scholarly content on social media.

Sugimoto and colleagues mention that even in the academic context, the definition and categorizations of social media platforms differ. Thus, these authors classified platforms according to their major functionalities, including social networking; social bookmarking and reference management; social data sharing; blogging;

microblogging; wikis; and social recommending, rating, and reviewing services (Sugimoto et al., 2016). In this chapter, to investigate users' engagement on social media with the KU pilot collection titles, a number of platforms are used and their data are compared with citations from a number of databases.

4.3.1 Facebook

According to the statistics portal Statista (http://www.statista.com), as of January 2018, Facebook was the most popular social network platform worldwide, with more than 2 billion users ("Leading global social networks 2018", 2018). The total number of Facebook users far exceeds that of Twitter users—who numbered around 330 million as of January 2018. However the number of users is not reflected on the number of mentions of journal articles. Xia and colleagues (2016), in their coverage analysis of articles from the general science journal *Nature*, found that there were more mentions of these articles on Twitter than on Facebook. This finding shows that Twitter is one of the preferred platforms for sharing research outputs.

In April 2015, following the introduction of Facebook's new v.2.0 API, the company disabled searches of public Facebook posts. Previously, this had been the content collection method most commonly used by social media monitoring tools (Meyers, 2014). Reseller companies began to sell public Facebook data, in particular relating to brand mentions on the platform. In the present study, Salesforce Radian6[2] software was used to collect public mentions of monograph titles on Facebook.

In order to extract information on Twitter and Facebook, the social studio dashboard of Radian6 (https://socialstudio.radian6.com/) was used. For each title, social media events were collected by search for the title name and author name strings on the dashboard.

4.3.1.1 Facebook findings

A total of 96 relevant Facebook posts were identified, in which 64 distinct Facebook users mentioned 21 of the KU pilot collection titles between the day the titles were

[2] Radian6 is a social media monitoring platform designed to help marketing professionals study customers' opinions of their products in real-time.

uploaded to OAPEN and 22 June 2017. The most frequently mentioned titles were "Human Rights and Democracy" (25 posts), "Constructing Muslims in France" (14 posts), "Biological Relatives" (nine posts), and "In Search of the Amazon" (nine posts) (Figure 4.1). These posts were mostly from publishers, Unglue.it repository users, and one author (Todd Landman, the author of "Human Rights and Democracy"). Most of these posts were categorized as link posts, which means that they contained a link.

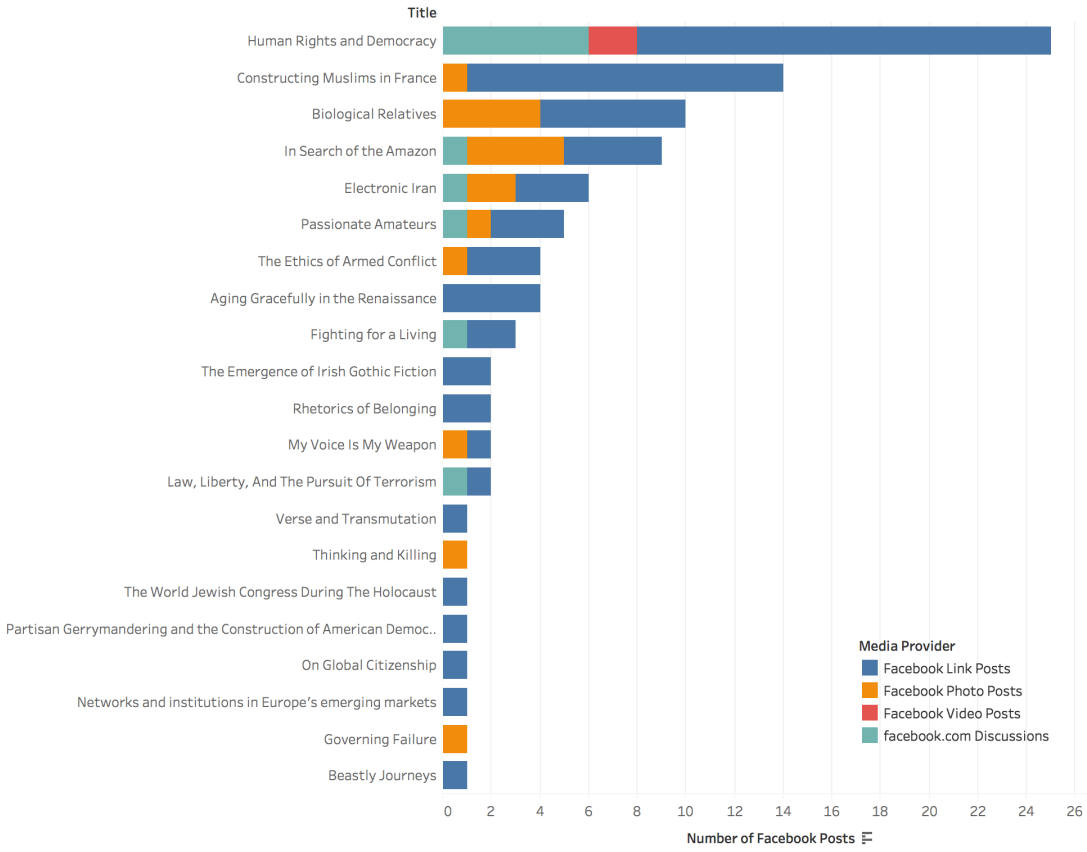

Figure 4.1: Number of posts mentioning each title on Facebook according to type of post

Besides those containing links, some Facebook posts contained photos (categorised as photo posts), which were mostly of the books' cover images. There were only two video posts, which were of Todd Landman's talks and were shared by the author himself. The discussion posts were mainly from the "Human Rights and Democracy" Facebook account, which is dedicated to Landman's book. Facebook was more actively used to promote Landman's book "Human Rights and Democracy" than other titles.

4.3.2 Twitter Mentions

The microblogging platform Twitter works differently from Facebook. Initially, Twitter allowed its users to post messages limited to 140 characters. In November 2017 this limit was doubled to 280 characters. On Twitter, users typically follow other users based on the content of their public tweets. As opposed to Facebook, where a user must approve access to their feed, on Twitter, by default, any user can view another user's public tweets without their approval. Another aspect of Twitter is that readers are not required to have a Twitter account to read tweets, which means that anyone on the Internet can discover and access tweets. These features make Twitter suitable for scholarly communication (Clarke, 2009).

Mahrt, Weller and Peters (2014) state that researchers, as well as institutions and research project coordinators, use Twitter to advertise their own research, events, publications, or other updates in much the same way that other commercial, political, or societal actors do in their marketing efforts on the platform (Kortelainen & Katvala, 2012; Sammer & Back, 2011). According to these authors, having a well-connected Twitter account and using pertinent hashtags helps increase the visibility of one's own research. Terras stated that tweets had helped disseminate and increase the downloads of her OA article (Terras, 2012). Thelwall and colleagues, in their research, found articles' coverage on Twitter was high. However, they found negative correlations between articles' Twitter mentions and their citations, due to the fact that more recent articles in their study was cited more but had no citations yet (Thelwall, Haustein, Larivière, & Sugimoto, 2013). Another study examined the effects of social media exposure for articles published in the *International Journal of Public Health* and found no effect of social media exposure on the number of downloads and citations, even though OA articles were downloaded more (Tonia, Van Oyen, Berger, Schindler, & Künzli, 2016). In this study, to measure the visibility of the KU pilot collection titles, tweets mentioning the titles were collected and correlations with various metrics, as well as with citations, were examined.

In addition to being used for advertising and the creation of personalized newsfeeds, users' Twitter data are also mined commercially for business insights. Bruns and Burgess (2016) state that in addition to commercially motivated developments in techniques for social media data analysis, both the social and behavioural sciences

and the digital humanities have been developing ever more sophisticated and large-scale methods for dealing with these data. Although they are often motivated by different questions, the commercial and academic fields rely on similar tools to access and analyse data, and thereby operate in ways that entangle scientific practice with the evolving markets in user data. For this reason, there are two different approaches on mining Twitter data: commercial, market-oriented research and scholarly, scientific research. Commercial and academic tools and methods can be used side-by-side in both areas. Commercial tools are useful for mining the Twitter platform when the aim is to access historical data, because Twitter's public API does not make available all of their data.

In this study, the commercial software Salesforce Radian6 social media cloud service, which was provided by the Curtin University School of Business, was used to collect social network data. The Salesforce Radian6 cloud service provides coverage of discussions on the social web, covering hundreds of millions of blogs, comments, publicly available Facebook posts, and all of Twitter's historical data. It is primarily used by businesses to monitor brand mentions across the social landscape.

There are also other free alternatives with which to collect and analyse Twitter data. One of these is TCAT, which is one of the tools provided by the Digital Methods Initiative (DMI). Located in Amsterdam, the DMI is an Internet studies research group, comprised of new media researchers and PhD candidates. They design methods and tools for repurposing online devices and platforms (such as Twitter, Facebook and Google) for research into social and political issues (The Digital Methods Initiative, n.d.). TCAT cannot access historical Twitter data, but can capture tweets in real time. The software has features such as the ability to capture tweets according to keywords entered by the user and to take a "1 percent" random sample of all tweets on Twitter (Borra & Rieder, 2014).

By visualizing TCAT data using the open-source network analysis and visualization sofware package Gephi (http://gephi.org), it is possible to locate and identify Twitter users. TCAT and Gephi make it possible to analyse a user's visibility according to the number of mentions they receive and their user statistics, including how many followers they have, who follows them, and how many favourites they have received

(The Digital Methods Initiative, 2015). The combination of these tools helps to visualize social networks in the Twittersphere and identify the groups and users that disseminate information of interest.

Since TCAT cannot access Twitter's historical data but collect through Twitter's API, it would have been better to track the KU pilot collection titles from their announcement dates, before publication. For this reason, it was not possible to track any of the titles in the KU pilot collection from the point of publication: all of the titles were published before this study commenced. However, as part of a project conducted for UCL Press by the KU Research team, mentions of Daniel Miller's "How the World Changed Social Media" title were tracked using TCAT for the period of 1 July 2016—five months after the book's publication in February 2016—until 13 April 2017. This allowed to capture tweets related to a title from its earlier period after its been published.

An important aspect of Twitter, as it relates to scholarly communication, is that it provides an environment that supports the formation of weak connections between users (Clarke, 2009). According to Granovetter's weak tie theory, innovations often travel most effectively via weak connections (Granovetter, 1973). He argues that typically, researchers are already familiar with the ideas and work of their immediate colleagues and friends, whilst a colleague with whom they communicate only occasionally, such as at conferences, is more likely to be a source of novel information. He states that such distant colleagues are more effective at spreading a researcher's novel ideas. This is because the researcher's close colleagues and friends are likely to know many of the same people the researcher does, whereas the distant colleague is more likely to have different people in their professional and social networks.

Applying this idea to the structure of Twitter's open, content-centric network, Clarke argues that the platform is good for information diffusion via weak ties. Thus, Twitter is a powerful tool for communicating scientific research, scholarship, and innovative ideas beyond one's immediate peer group. Clarke argues further that although 140-character posts are not a substitute for other forms of formal and informal communication, Twitter augments such communication channels, increasing their impact and reach in ways that other networks cannot (Clarke, 2009).

In line with this idea, Twitter data and the TCAT and Gephi tools were used to visualize a network structure for the diffusion of Daniel Miller's title on the Internet and to identify the node that disseminated this information to other groups. A network structure is formed on Twitter when connections (links, edges, and ties) are created among Twitter users (nodes) by tweeting, and retweeting on a specific title.

4.3.2.1 Twitter findings

Using the Radian6 service, 493 tweets from 309 different authors were identified which mentioned KU pilot collection titles between the day they were uploaded and 22 June 2017. There were more tweets than Facebook posts mentioning these titles, which is in accordance with the findings of Xia and colleagues (Xia et al., 2016). In addition, the titles' rankings according to the number of mentions they received on Twitter differed from those on Facebook. The three most mentioned titles were "The Ethics of Armed Conflict" (46 mentions), "Passionate Amateurs" (36 mentions), and "Beastly Journeys" (33 mentions), and the least mentioned title was "Governing Failure" (1 mention). Figure 4.2 shows the number of mentions for each title. Most mentions came from the official KU Twitter account (30 tweets), followed by the Unglue.it account (21 tweets).

Figure 4.3 shows the numbers of tweets mentioning a title with respect to the period that had elapsed since the title was made OA. All 28 titles were made OA for the first time when uploaded to the OAPEN repository. Half of the tweets were produced in the first four months, with 35.35% produced in the first two months. In Figure 4.3, a plateau in tweet numbers can be seen after the first 850 days following publication. Eysenbach refers to the period from the first tweet until the plateau (the first 850 days) the "network propagation phase", during which the new information is propagated through the Twitter network. He refers to the period following this phase as the "sporadic tweetation phase", where mentions only occur sporadically.

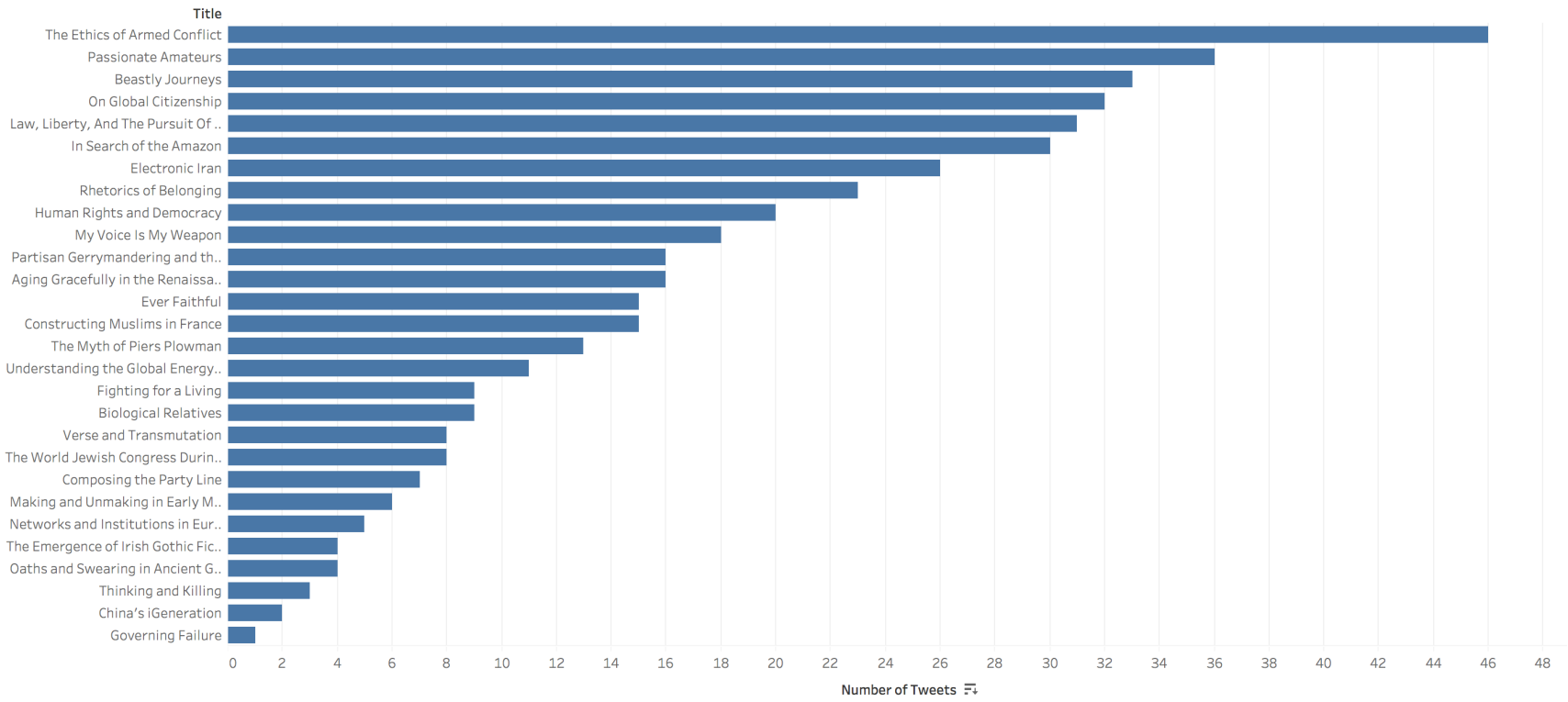

Figure 4.2: Number of tweets mentioning each KU pilot collection title.

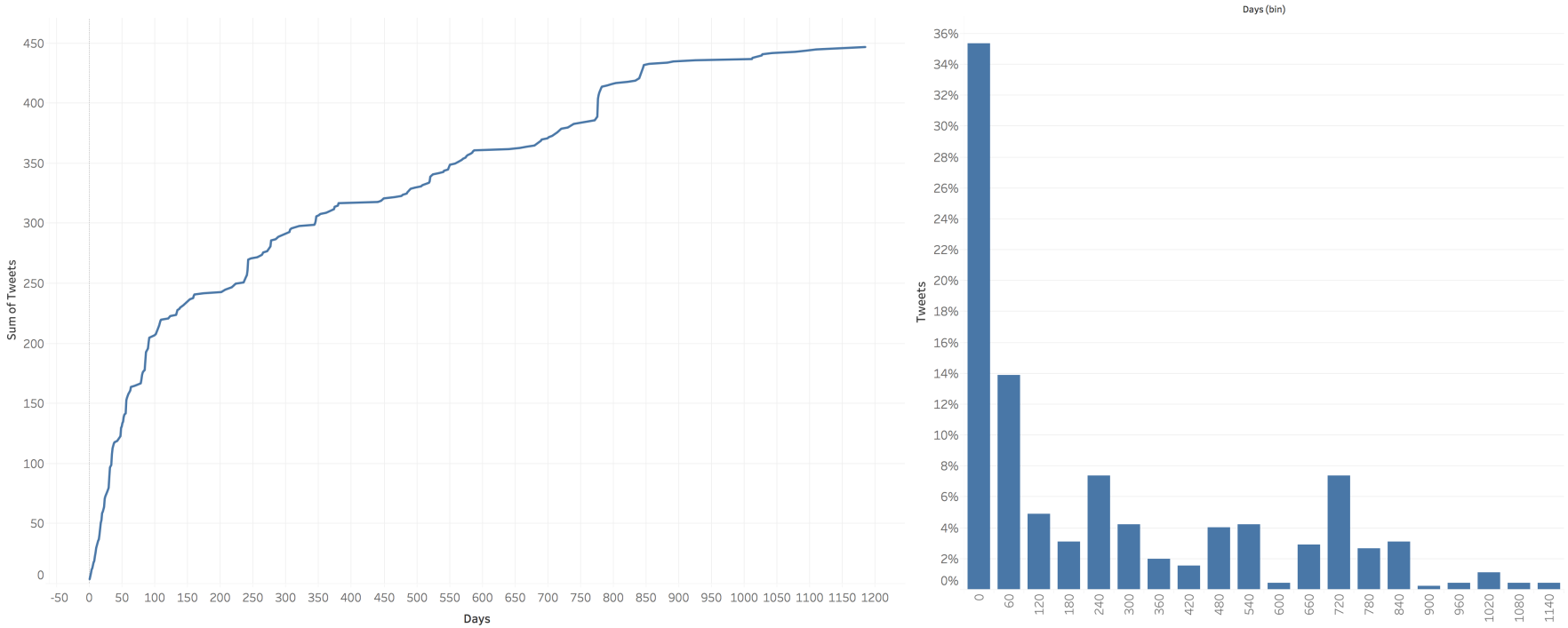

Figure 4.3: Number of tweets mentioning the KU titles on a time scale. Most mentions occur in the first four months following publication.

Because this study was started two years after the KU pilot collection was made OA, it was not possible to track Twitter mentions of the titles by using the TCAT software to graph the network structure. Instead, as a case study, tweets mentioning the title “How the World Changed Social Media” were tracked for nine months (from 1 July 2016 until 1 April 2017) as part of a study involving titles from the OA monograph publisher UCL Press (Montgomery et al., 2018). Tweets for this title were captured five months after the title’s publication. The title was mentioned in 181 tweets from 103 distinct users, and 82.3% of these tweets contained links. During the network analysis of these tweets and their users, four groups within the network were investigated. The network had a total of 109 nodes (or Twitter accounts) and 191 edges. Among these four groups, the biggest group is shown in purple and contains the UCL Press (‘uclpress’), ‘UCL Why We Post’ book series (‘UCLWhyWePost’), and the book’s author Daniel Miller’s (‘dannyauth’) account; the green group contains the account of openlibra (‘openlibra’), a website hosting free books and four other Twitter accounts retweeting its tweet about the book; the blue group contains 'bokofil' account mentioning amazon Twitter account; and the orange group contains the account of Ritu Gairola ('ritu_gairola'), an assistant professor of cultural antrhopology, whose tweet mentioning the book was retweeted by another account ('betoceforpeople').

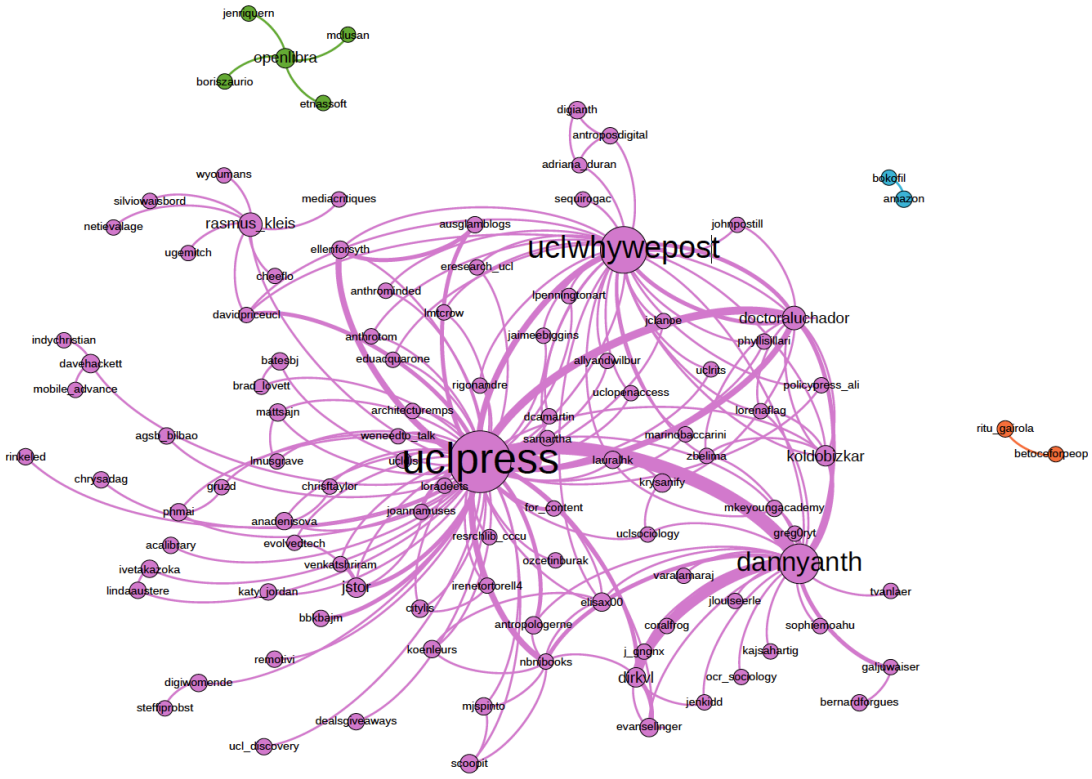

Figure 4.4: Twitter network graph for the title "How the World Changed Social Media" published by UCL Press.

In Figure 4.4, the edges are rotated clockwise from the account mentioning to the mentioned account, or from retweeting account to the tweeting account. The size of a node is proportional to the node's number of mentions it receives.

Although this graph shows diffusion of the title, it does not show Twitter accounts following other accounts. This graph therefore focusses on mentions but does not show the overall reach of those that might have seen information on the title. The full visibility graph cannot be reconstructed because the state of the follower network at the time of mentions is no longer available.

Alperin and colleagues (2019) in their studies, which includes follower accounts, analysed diffusion patterns of research articles on Twitter from a sample of 11 articles from two access biology journals that were shared on Twitter at least 50 times. They found that the diffusion pattern can take different forms, and most of the articles were shared within a single community with limited diffusion to the public. The PageRank algorithm can be used to identify important nodes. PageRank is one of the algorithms that Google uses to rank web pages in their search results. It is

named after Larry Page, one of Google's co-founders. The algorithm works by counting the number and quality of links to a node to determine a rough estimate of the node's importance (Google, 2011). However, important nodes do not necessarily receive the most mentions. It can also be a node being mentioned by most mentioned nodes as well, thus having high quality of links. In other words, it can be a Twitter user not having too many mentions but being mentioned by some popular users having many mentions, which makes them an important user according to this algorithm. Therefore these important nodes are the 'uclpress', 'dannyanth', 'uclwhywepost', 'doctoraluchador', 'koldobizkar', and 'rasmus_kleis' Twitter accounts.

In addition to the use of the PageRank algorithm, a betweenness centrality analysis was also conducted to identify the nodes' centrality in the network. Betweenness centrality is equal to the number of shortest paths from all nodes to all others that pass through a node. This node can be seen as a bridging node for reaching other nodes. The bridging nodes for "How the World Changed Social Media" are 'uclpress', 'uclwhywepost', 'doctoraluchador', 'elisax00', 'ellenforsyth', and 'lauralhk'.

This shows the nodes that are responsible for pushing or receiving information from different areas of the larger network. These nodes are more engaged in information sharing. In other words, these are the Twitter accounts that are influential in the dissemination of information. Betweenness centrality analysis is helpful in identifying Twitter accounts that are key for the titles' dissemination across the Twittersphere. Therefore, the Twitter accounts that are identified using the PageRank algorithm and betweenness centrality analysis are crucial for publishers and repositories' promotion of their books.

4.3.3 Blogs

In the late 1990s, the availability of free, easy-to-use blog publishing and hosting software and services lowered the barriers for participation in online publishing (Fox & Lenhart, 2006). Following this development, researchers began to blog in order to share their ideas and research (Kjellberg, 2010). There is no precise distinction between a blog and a scholarly blog. Scholarly blogs have most commonly been

defined in accordance with Puschmann and Mahrt's (2012) loose definition of "blogs written by academic experts that are dedicated in large part to scientific content". Science-only aggregators, including ResearchBlogging.org (Shema et al., 2014), a platform that was in operation until April 2017, direct readers to blog posts that refer specifically to peer-reviewed research. Bloggers who discussed peer-reviewed research would register their blog as part of ResearchBlogging. The site's human editors would then examine blogs to ensure that they followed ResearchBlogging's guidelines and were of appropriate quality.

Few altmetrics studies have been conducted on scholarly blogging (Priem, 2014). Most of the research on blogs investigates the coverage of research articles and the connection between citations in blogs and in formal scholarly publications. Shema and colleagues argue that although there is evidence that mentions of scholarly articles in blogs correlates positively with citations, only a small percentage of articles are covered on blogs (Shema et al., 2014). The highest coverage was for PLOS articles, 7.5% of which were mentioned on blogs (Priem et al., 2012). Costas, Zahedi and Wouters (2015a) found 1.9% of the 500,229 articles they examined were mentioned on blogs. Unfortunately, in this study, none of the 28 monograph titles were cited in blog posts collected from ResearchBlogging.

4.3.4 Wikipedia

Wikipedia is a platform with wide use among scholars and non-scholars. Whether Wikipedia shapes academic production is a controversial topic. Its usage shows that it is one of the most consulted reference site on the Internet. After Wikipedia introduced structured scientific citation use, confidence to it as an information organiser increased among users. Wikipedia also provides valuable background reading for researchers (Nielsen, 2007). Because of its popular coverage of topics, Wikipedia is an important intermediary for the diffusion of science to a broad audience (Teplitskiy, Lu, & Duede, 2017). On Wikipedia, authors reference OA journals more than paywalled journals. Teplitskiy and colleagues also argue that although Wikipedia authors prefer to refer to OA literature, they prioritize references to high-impact journals, regardless of whether they are open or closed access.

Shuai and colleagues found a positive correlation between mentions on Wikipedia and citations in formal scholarly publications (Shuai et al., 2012). They showed that papers, authors, and topics that are mentioned on Wikipedia have more citations than those that are not. However, Marashi and colleagues showed that the inclusion of scholarly references in Wikipedia does not affect the citation "propensity" of these articles (Marashi et al., 2013). In this study, to examine whether the 28 KU titles were referred to in Wikipedia articles, the following query was entered into the Google search engine:

```
"Title" + site:wikipedia.org
```

For each title the number of unique Wikipedia pages that were returned was recorded.

4.3.4.1 Wikipedia findings

A total of 23 Wikipedia articles, which referred to 13 different monographs, were identified. The most-cited monograph was "Oaths and Swearing in Ancient Greece", which was referred to in four articles. Four other titles were each cited in two articles. Nine of the 23 Wikipedia articles contained no link to the PDF file of the titles. It is possible that the authors of these Wikipedia articles did not know that these books were OA.

In six of the Wikipedia articles, the references to the KU titles directed users to Google Books, and only one of these six articles directed users to the freely accessible content in Google Play. The remaining five articles directed users to Google Books pages, which contained only a presentation page of the title, or to Google pages where users had to pay for the content. As has been noted, most of the publishers did not provide the free PDF content on Google Books. There were four articles on the French version of Wikipedia referring to "Beastly Journeys", which each directed users to the free version of the book on The Internet Archive. Only two of the 23 Wikipedia articles directed users to the OAPEN repository for the OA PDF of the book.

The language distribution of the relevant Wikipedia articles was as follows:

Table 4.2: Language distribution of articles citing KU pilot collection monographs on Wikipedia.

Wikipedia language	Number of Articles in Wikipedia	Number of Distinct Monograph Titles
English	14	10
French	6	2
German	2	2
Catalan	1	1

One of the issues with the Wikipedia articles was that the reference sections provided a link to the title's ISBN number, which directed users to platforms including Google Books, Amazon, and Open Library. Google Books and Amazon do not provide free access to most of them. The Open Library project (OpenLibrary.org), which is part of The Internet Archive (archive.org), does not host KU pilot collection titles. However, The Internet Archive's website does host them. This is not mentioned on Wikipedia, and it is not possible to reach the titles hosted in The Internet Archive through Open Library. The difference between these two sites is that Open Library is a catalogue of books with the mission of offering "One web page for every book". In contrast, archive.org offers free public access to all sorts of materials, including books, music, video, web pages, and software, which users upload (Kahle, 2018). Therefore, Wikipedia should include links to repositories such as OAPEN, HathiTrust, and The Internet Archive in their articles' reference sections. In this way, it can provide freely accessible monographs to its readers.

4.3.5 Amazon, Goodreads, and Google Books

In order to obtain reviews and ratings of the 28 KU pilot collection titles, three platforms were investigated, including Amazon, Google Books and Goodreads. Although these three platforms operate in different ways, they all offer book reviews and ratings. These reviews and ratings can provide insight into these monographs' reception amongst the broader public.

According to Wu and Zheng (2012), reviews affect book sales and tend to be positive. However, the impact of one-star reviews is apparently greater than that of five-star reviews (Chevalier & Mayzlin, 2006). Forman and colleagues found that in general, extreme ratings, such as 1 or 5 stars, are more helpful for users, based on user feedback (Forman, Ghose, & Wiesenfeld, 2008). The Amazon platform shows

whether a reviewer of a book has purchased the book and also displays the format of the book that was purchased, such as hardcover, paperback or Kindle edition. Given a sufficient number of reviews, this feature can provide insight into the format readers prefer to buy.

Kousha and Thelwall (2016) compared number of reviews on Amazon with citation counts from the Thomson Reuters Book Citation Index (BKCI) for 2,739 books indexed across various disciplines. They found statistically significant but weak correlations between them. They also found that many high-impact BKCI scientific books had no reviews on Amazon, but most of the bestselling textbooks were cited. Nevertheless, the number of Amazon reviews reflects the coverage of a book's readership to a much greater extent than the number of citations does.

Goodreads, which was launched in 2007 and acquired by Amazon in 2013, claims to be the world's largest book recommendation site for readers, with 68 million reviews across 2 billion titles as of March 2018 (Goodreads, n.d.). According to Jordan Weissman, book industry research from the Codex Group estimates that 46% of recommendations are made by 11% of book buyers (Weissmann, 2013). Weissman reports that roughly 29% of Goodreads users had learned about the last book they bought either from the site or from another book-focused social network. Weissman states that this rate of learning from traditional social network platforms, such as Facebook, is 2.4 percent. This shows how effective Goodreads is than traditional social network platforms for book recommendation. Thus, the purchase of Goodreads allowed Amazon to learn about readers' thoughts and habits. Therefore, Goodreads review counts presents itself as a good proxy candidate for readership. In their study, Zuccala and colleagues found a weak correlation (0.212) between citation counts and Goodreads rating counts for history books (Zuccala, Verleysen, Cornacchia, & Engels, 2015).

Google Books is a service that offers full-text book searches for titles Google has scanned, converted to text, and stored in its digital database. Currently, books are provided either by publishers and authors or by libraries through Google Books library projects. As of 2015, Google Books had scanned 30 million volumes (Wu, 2015). For books still in print or ebooks on sale, the site offer links to the publisher's and booksellers' sites. They also allow registered users with Google accounts to post

reviews and ratings for books. Kousha and colleagues examined citations in Google Books, Google Scholar, and Scopus for a sample of 1000 books submitted to the 2008 U.K. Research Assessment Exercise (Kousha, Thelwall, & Rezaie, 2011). They found that Google Books and Google Scholar citations to these 1000 books were 1.4 and 3.2 times more common than were Scopus citations.

4.3.5.1 Google Books, Amazon, Goodreads findings

Although all the KU pilot collection titles were registered on Google Books, only one book review (for the title "Ever Faithful") was found on the platform. Access to the content on Google Books was also investigated. In spite of all the books being OA, only three monographs ("Thinking and Killing", "Oaths and Swearing in Ancient Greece" and "The World Jewish Congress During the Holocaust") published by De Gruyter, had free content access from Google Books. Two Google Books pages were found for "Oaths and Swearing in Ancient Greece". One page directed users to the free content, while the other page asked US$84.02 for the same content. There were no e-book versions of the 15 KU pilot collection titles with Google Books. It was also found that Google Books was selling 11 titles through its Google Play platform.

Only 20 of the titles' e-book versions were available on the book-selling platform Amazon, and only two of them (published by De Gruyter) had free access. E-book prices ranged from US$11.14 to US$89.14, as of March 2018. Eight titles did not have their electronic versions available on Amazon. Only four five-star reviews were found for three different titles. These reviews were given for books ranging in price from US$11.14 to US$16.68. Only one reviewer was identified as a verified customer, and they had purchased the electronic Kindle version of a book.

The book review site Goodreads was queried using the 'Rgoodreads' package in R, which connects to the Goodreads API (https://www.goodreads.com/api). All of the 28 KU pilot collection titles were registered on the portal. Using the ISBN identifiers obtained from the OAPEN metadata, the site was queried for reviews of these titles. Forty-nine ratings were given for 16 different titles. Among these, four reviews were written for four different titles. The most rated book, with 13 ratings, was "Biological Relatives". The second most frequently rated titles were "In Search

of the Amazon" and "On Global Citizenship", with five ratings each. It was not possible to deduce readers' preferred format of these books, because only seven titles were registered as ebooks. The rest were registered as print books.

4.3.6 Mendeley reference manager

Reference managers such as Mendeley, CiteULike and Zotero allow users to save references into online referencing libraries or to share them in groups. This places Mendeley in a broader category referred to as "scholarly social bookmarking systems". Reference managers are an important source for altmetrics research because they offer data on scholars' libraries. These platforms make it possible to count resource-specific bookmarking actions, including the number of users saving a particular resource, or the distribution of bookmarks based on users' academic status. This study focussed on the Mendeley platform because of its wider coverage of scholarly outputs and the availability of data for free.

Mendeley is a desktop software application and also a web platform. Its web platform is an online social networking and collaboration medium for researchers. Launched in 2008, the company was acquired by Elsevier in 2013. As of March 2012, they also stated that the platform covered more than 34 million papers (Haustein et al., 2014a). In September 2013, Elsevier announced that the Mendeley platform had registered 2.5 million users all over the world. Its popularity has made Mendeley one of the most studied reference managers.

Several research studies have been conducted on Mendeley. In a number of these studies, the Mendeley findings were compared with those from citation databases, including WoS and Scopus. In a study conducted by Costas and colleagues, which involved 500,216 WoS publications (both articles and reviews) with DOIs between July and December 2011, Mendeley was found to be the social media source most similar to citations in terms of their distribution across fields of science (Costas, Zahedi, & Wouters, 2015b). Another study by Bar-Ilan and colleagues sampled 1,136 unique papers authored by 57 presenters who attended the 2010 Leiden Science and Technology Indicators (STI) conference (Haustein et al., 2014a). Of these 1,136 documents indexed in Scopus, they found 928 with at least one Mendeley bookmark (82%) and 961 (85%) with at least one citation in Scopus. Bar-

Ilan et al. also found a significant correlation (r=0.45) between an article's number of bookmarks in Mendeley and number of citations in Scopus. In another study, which involved 1,613 papers published in *Nature* and *Science* in 2007 (Li, Thelwall, & Giustini, 2011), positive correlations (r=0.60 and r=0.54, respectively) were found between the articles' bookmark counts in Mendeley and their citation counts in WoS. Although bookmarks in Mendeley are significantly correlated with the number of citations, a study conducted by Mas-Bleda and colleagues, which examined the 250 most cited researchers in Europe from 1981 – 2008 in each of 21 disciplines, found that very few of these researchers had Mendeley profiles (Mas-Bleda, Thelwall, Kousha, & Aguillo, 2014).

Haustein and Larivière (2014), in their work on readership counts on Mendeley, examined 1.2 million documents published in journals from four disciplines (biomedical research, clinical medicine, health and psychology). They found that approximately two thirds of these documents were bookmarked by at least one user on Mendeley. The majority of these users were PhD students, postgraduate students and postdoctoral researchers.

An important limitation of all of these studies is that even if Mendeley users bookmark a title, it does not mean that they have read these titles. In Mohammadi and colleagues' survey of 860 Mendeley users, 55% had read or intended to read at least half of their bookmarked publications. Approximately 85% of users bookmarked publications in order to cite them, 50% bookmarked publications for professional use, 25% bookmarked publications for teaching, and 13% bookmarked publications for educational activities such as assignments (Mohammadi, Thelwall, & Kousha, 2015). In the current study, the number of users who had bookmarked KU pilot collection titles, along with these users' academic statuses, were determined and examined.

4.3.6.1 Mendeley reference manager findings

In order to identify users who had bookmarked the KU pilot collection titles on the reference manager platform Mendeley, the platform's API (https://api.mendeley.com/) was queried with these titles' ISBNs using an R script. Query results contained the title, the academic status of the reader, and the number

of readers for each academic status. The results showed that 28 different entries had been made for 20 titles. Twenty-six of these were for the entire book, and two were for a chapter of the book. These 28 different entries had a total of 288 readers. The most bookmarked title was “Biological Relatives”, with 74 readers; followed by “My Voice is My Weapon” and “On Global Citizenship” with 21 readers each (Figure 4.5). Most bookmarks were made by graduate students (159), followed by undergraduate students (42). The number of researchers, professors, and lecturers bookmarking these titles was small compared to the number of students, as shown in Figure 4.6.

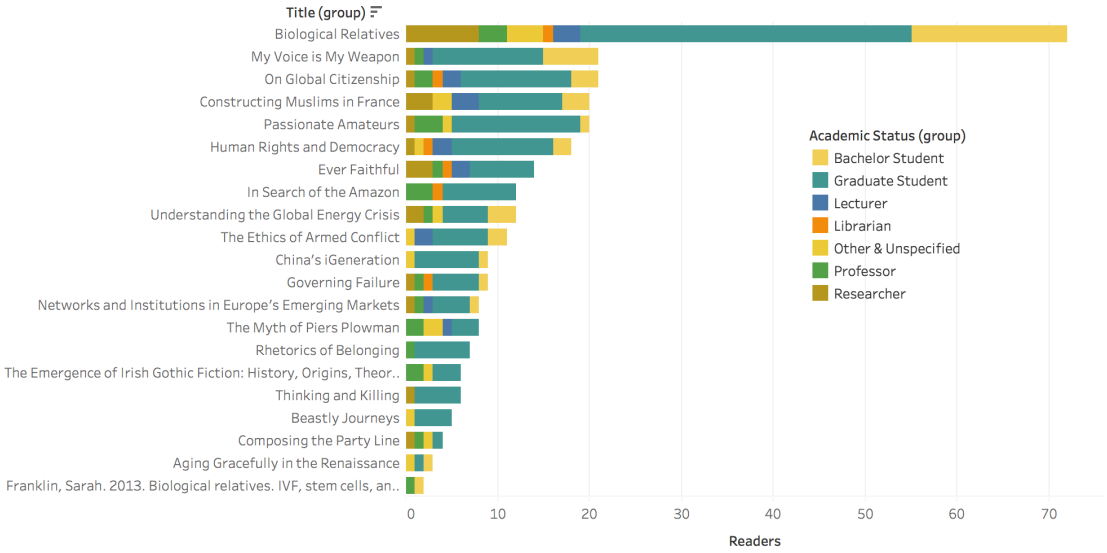

Figure 4.5: Distribution of titles’ bookmarks according to readers’ academic status.

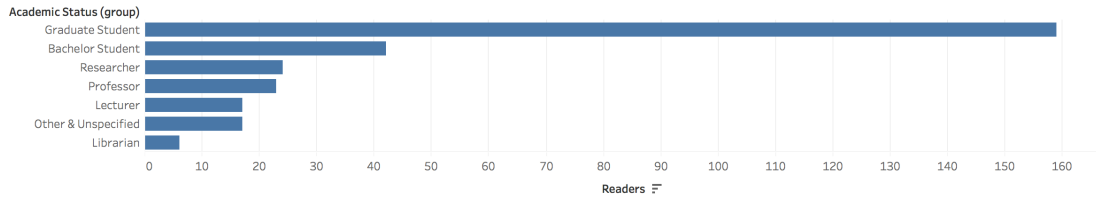

Figure 4.6: Academic status of Mendeley readers of the full set of KU pilot collection titles.

4.3.7 Annotation platforms: Hypothes.is and PaperHive

A web annotation is an online annotation made on a web resource, such as a web page or a PDF document. Users can write, modify or remove these notes on the web resource without changing the resource itself. The annotations act as a new layer on

top of the existing resource. Because the web page and its annotations usually reside on different servers, only users of the annotation system can view or manipulate annotations on a web resource.

Several online annotation services exist. This study focuses on the Hypothes.is and PaperHive platforms. Founded in 2011, Hypothes.is is a non-profit annotation service. In December 2015, Hypothes.is was a founding member of a coalition of scholarly publishers, platforms, libraries, and technology organizations with the goal of creating an open, interoperable annotation layer over their content.

Hypothes.is can be used with a browser plug-in or directly via its own page by entering the address. Annotations can also be used locally. For example, using a web browser, it is possible to open a PDF file, whether it is online or local, and annotate it using the browser plug-in. The browser plug-in shows other users' comments on the PDF document, because the annotation system recognizes each PDF. In this study, this feature makes it possible to see the annotations for a single title, even when the file resides in different repositories. The only condition is that these PDF files are identical or derivatives of the PDF file that resides on the OAPEN repository, so that their fingerprint ID remains the same. Another feature of Hypothes.is is that annotations can be either private or public. Public annotations thus allow us to examine the number of readers interacting with these titles and to investigate which parts of these books attract the most interest.

Only titles registered on the PaperHive site can be annotated on the site. The platform first downloads the PDF file for the reader to read and annotate. In February 2018, PaperHive contained 1.2 million academic articles and books from various publishers, which can be read and discussed in real time (Sharma, 2018).

4.3.7.1 Annotation platforms findings

To investigate annotations of the 28 KU pilot collection titles on the annotation platform Hypothes.is, all of the titles' PDF files were downloaded. Because Hypothes.is supports local usage of PDF files, all files were examined manually. None of the titles had any annotations on the Hypothes.is platform.

The annotation platform PaperHive was also examined, and only one discussion, started in 2016 and related to the title "On Global Citizenship", was found. Since

PaperHive does not offer the ability to work on local PDF files, their site was used to determine whether there were any annotations or notes on the KU pilot collection titles.

4.3.8 Citation Databases: Scopus, WoS and Google Scholar

Researchers consult citation databases to track the most-cited articles on a particular topic. For many years, the main source of citation data was the Thomson Reuters' Web of Knowledge database, now called WoS. In 2004, two new service emerged: Elsevier's Scopus and Google's Google Scholar. Studies have compared the coverage of Scopus, WoS, and Google Scholar according to language and scientific discipline (Harzing & Alakangas, 2016; Waltman, 2016). In August 2017, Elsevier announced that Scopus had over 69 million records, including more than 21,950 peer-reviewed journals, eight million conference papers, and more than 150,000 books, with another 20,000 added each year (Elsevier, 2016). Today, WoS belongs to Clarivate Analytics, and as of 18 February 2018, it covered over 69 million records, including 18,200 journals, 88,000 books and 10.2 million conference papers (Clarivate, 2018). Google, on the other hand, has not published the size of its Google Scholar database. However, a study published in 2015 estimated that it contained roughly 160 million documents in May 2014 (Orduña-Malea, Ayllón, Martɪn-Martɪn, & López-Cózar, 2015). Although Google Scholar is free and has more coverage, its coverage is not as transparent as the other two common citation indexes, Scopus and Web of Science. Google Scholar does not descripbe what criteria they are using to select "scholarly" material or the set of objects that re indexed. There are also non-scientific documents which are occasionally covered and some duplicate entries for the same research output (Delgado López-Cózar, Orduña-Malea, & Martín-Martín, 2019).

Studies on altmetrics typically compare social media metrics with citations using Scopus, WoS, and Google Scholar. Eysenbach examined whether social metrics could be a predictor of citations, and investigated tweets that contain links to articles in the Journal of Medicinal Internet Research. He found that tweets within the first three days of these articles' publication can predict highly cited articles (Eysenbach, 2011). Although these studies show promising results for journal articles, there is

unfortunately an insufficient number of studies on the relationship between these metrics and OA monographs' citations.

One of the few studies conducted on monographs, Snijder (2016), found a small positive effect of OA publishing on citation scores, and only a weak relationship between the number of tweets mentioning a book and the number of citation it gets. According to Snijder, it is probable that the factors that affect the number of book citations do not significantly affect the number of tweets that mention a book. The current study first examines the visibility of the 28 KU pilot collection monographs on these citation platforms and subsequently investigates whether any correlations exist with the social media metrics that have been collected.

4.3.8.1 Citation Databases findings

Coverage of the KU pilot collection titles was examined in three citation databases. The titles were queried in the Scopus book title list of December 2017, which contained 158,664 indexed titles. Only eight titles were found to be indexed by Scopus. Six out of the eight titles were found to have citations. The two most cited titles were "Governing Failure" (25 citations) and "Passionate Amateurs" (24 citations). Google Scholar indexed all 28 of the titles. The title with the most citations was "Biological Relatives" (162 citations), followed by "Passionate Amateurs" (65 citations). Only four authors had Google Scholar pages: Lawrence Warner ("The Myth of Piers Plowman: Constructing a Medieval Literary Archive"), Sarah Franklin ("Biological Relatives: IVF, Stem Cells and the Future of Kinship"), Niki Akhavan ("Electronic Iran: The Cultural Politics of an Online Evolution"), and Nicholas Ridout ("Passionate Amateurs: Theatre, Communism and Love").

WoS indexed all 28 titles except "Law, Liberty, and the Pursuit of Terrorism". The most cited monograph was "Biological Relatives" (69 citations), followed by "Passionate Amateurs" and "On Global Citizenship" (both with 28 citations). The least cited title was "The World Jewish Congress During the Holocaust", with only one citation (Table 4.3).

Table 4.3: Citations from Scopus, Google Scholar and WoS, sorted according to WoS citations.

Title	Scopus	Google Scholar	WoS
Biological Relatives		162	69
Passionate Amateurs	24	65	28
On Global Citizenship		48	28
China's iGeneration		43	21
Governing Failure	25	60	21
Fighting for a Living		32	19
My Voice Is My Weapon		46	16
Ever Faithful		18	14
Constructing Muslims in France		42	12
Electronic Iran		33	10
Partisan Gerrymandering and the Construction of American Democracy	6	18	8
Understanding the Global Energy Crisis		24	8
Human Rights and Democracy		31	8
Aging Gracefully in the Renaissance		7	7
Verse and Transmutation		8	6
The Myth of Piers Plowman		7	5
Rhetorics of Belonging		25	5
Making and Unmaking in Early Modern English Drama		8	5
Composing the Party Line		10	5
Networks and Institutions in Europe's Emerging Markets	4	15	5
The Ethics of Armed Conflict	5	11	4
The Emergence of Irish Gothic Fiction	4	13	3
Beastly Journeys		0	3
Thinking and Killing	0	3	2
In Search of the Amazon		8	2
Oaths and Swearing in Ancient Greece		14	2
The World Jewish Congress During The Holocaust		4	1
Law, Liberty, And The Pursuit Of Terrorism	0	5	

A correlation analysis conducted on the three citation databases for these 28 titles showed that the citation numbers within the three databases were highly correlated

with one another (Table 4.4). However, as there was only eight data points from Scopus, the statistical power of the comparison is questionable. Google Scholar and WoS were chosen as a proxy for the citation databases because of the number of titles they covered.

Table 4.4: Correlations between citation numbers within three databases.

Citation database	Correlation result
Google Scholar – Scopus	0.99
WoS – Scopus	0.97
WoS – Google Scholar	0.96

According to Harzing (2017), WoS's journal listing is considerably less complete in the HSS compared to the sciences. It also has limited coverage of non-journal publications. This is because WoS only includes citations from WoS-listed journals. Consequently, a vast majority of HSS publications and citations are ignored.

One of the advantages of WoS and Scopus is that it provides citations by year. The yearly citations for all 28 KU pilot collection titles are plotted in Figure 4.7. Some monographs were cited immediately after their publication. In addition, some titles were cited before their publication. The aggregated number of citations for each year from the beginning of 2015 until the end of 2017 ranged between 86 and 98. The number of titles covered for these years was between 21 and 23. No citation pattern over time could be identified for each title.

Figure 4.7: Yearly aggregated citations for the 28 KU pilot collection titles from WoS. "Biological Relatives" (in orange) has the most citations.

Table 4.5 summarizes the data collected from all the social media sources and citation databases for each title (excluding the case study title).

Table 4.5: Summary of datasets obtained from social media and citation database platforms for each title.

Title	Twitter Mentions (Mar 2014–Jun 2017)	Facebook Mentions (Mar 2014–Jun 2017)	Number of Articles Mentioning Title on Wikipedia	Number of Mendeley Readers	Number of Citations on Scopus	Number of Citations on Google Scholar	Number of Citations on WoS	Goodreads Ratings	Goodreads Rating Count
Aging Gracefully in the Renaissance	18	4	1	3		7	7	–	0
Beastly Journeys	35	1	4	5		0	3	3	1
Biological Relatives	17	10	1	74		162	69	3.46	13
China's iGeneration	2	0	1	9		43	21	3	1
Composing the Party Line	7	0	0	4		10	5	4	2
Constructing Muslims in France	16	14	0	20		42	12	4	3
Electronic Iran	28	6	0	0		33	10	3	1
Ever Faithful	18	0	0	14		18	14	3.75	4
Fighting for a Living	10	3	1	0		32	19	4	1
Governing Failure	1	1	0	9	25	60	21	–	0
Human Rights and Democracy	21	25	0	18		31	8	–	0
In Search of the Amazon	30	9	2	12		8	2	3	5
Law, Liberty, And The Pursuit Of Terrorism	34	2	2	0	0	5		–	0
Making and Unmaking in Early Modern English Drama	6	0	0	0		8	5	–	0

Title	Twitter Mentions (Mar 2014–Jun 2017)	Facebook Mentions (Mar 2014–Jun 2017)	Number of Articles Mentioning Title on Wikipedia	Number of Mendeley Readers	Number of Citations on Scopus	Number of Citations on Google Scholar	Number of Citations on WoS	Goodreads Ratings	Goodreads Rating Count
My Voice Is My Weapon	20	2	0	21		46	16	–	0
Networks and Institutions in Europe's Emerging Markets	5	1	1	8	4	15	5	–	0
Oaths and Swearing in Ancient Greece	4	0	4	0		14	2	3	1
On Global Citizenship	38	1	2	21		48	28	4	5
Partisan Gerrymandering and the Construction of American Democracy	16	1	0	0	6	18	8	–	0
Passionate Amateurs	45	5	0	20	24	65	28	3.5	4
Rhetorics of Belonging	25	2	0	7		25	5	4	2
The Emergence of Irish Gothic Fiction	4	2	2	6	4	13	3	4.67	3
The Ethics of Armed Conflict	46	4	0	11	5	11	4	–	0
The Myth of Piers Plowman	16	0	1	8		7	5	–	0
The World Jewish Congress During The Holocaust	9	1	0	0		4	1	–	0
Thinking and Killing	3	1	0	6	0	3	2	5	2
Understanding the Global Energy Crisis	11	0	1	12		24	8	–	0

Title	Twitter Mentions (Mar 2014–Jun 2017)	Facebook Mentions (Mar 2014–Jun 2017)	Number of Articles Mentioning Title on Wikipedia	Number of Mendeley Readers	Number of Citations on Scopus	Number of Citations on Google Scholar	Number of Citations on WoS	Goodreads Ratings	Goodreads Rating Count
Verse and Transmutation	8	1	0	0		8	6	5	1
Total	**493**	**96**	**23**	**288**	**68**	**760**	**317**	**Average: 3.77**	**49**

4.4 Correlation Analysis

A Spearman correlation analysis was conducted to examine the relationships between the following metrics: total downloads of the titles from the OAPEN platform, number of domains mentioning these titles, number of citations from WoS and Google Scholar, number of citations on Wikipedia, number of mentions on Twitter and Facebook, number of bookmarks on Mendeley, and number of ratings on Goodreads. The results of this correlation analysis are shown in Figure 4.8. In the figure negative correlations are in blue colour and positive correlations in red.

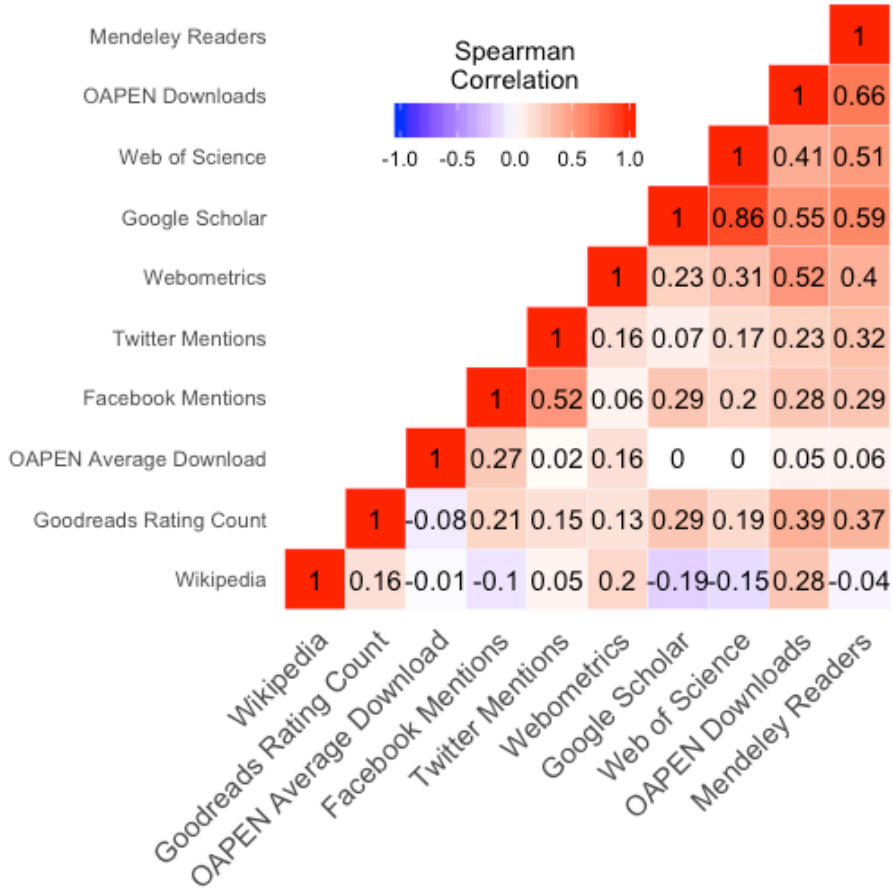

Figure 4.8: Correlation analysis results using heatmap.

Figure 4.8 shows that the highest correlation (0.86) is between Google Scholar citations and WoS citations. The second highest correlations is between Mendeley bookmarks and OAPEN Downloads (0.66), and the third highest between Mendeley bookmarks and Google Scholar citations (0.59). The positive correlation between Mendeley bookmarks and citations from citation databases is not surprising, since researchers, in particular graduate students, bookmark titles in order to cite them. Mendeley bookmarks are also moderately correlated with the number of domains

mentioning a title, as well as with number of OAPEN downloads and Facebook mentions.

Mentions on social networking platforms do not seem to correlate strongly with other metrics. Twitter mentions have a weaker correlation with other metrics than Facebook mentions. In addition, the number of citations on Wikipedia and the average monthly downloads from the OAPEN platform are not strongly correlated with other metrics.

The p-values of the following correlations were above 0.05: OAPEN average downloads, Twitter, Wikipedia, and Facebook with all data sources. The only exception for Facebook was the correlation between Facebook mentions and Twitter mentions, which had a correlation coefficient of 0.52 with p less than 0.01. Only p-values below 0.05 were accepted as significant for the correlations between Mendeley bookmarks, Goodreads ratings, OAPEN downloads, number of domains mentioning titles, and citation numbers from Google Scholar and WoS.

To summarize: strong correlations are observed between citations on Google Scholar, citations on WoS, and Mendeley bookmarks. Among these metrics, Mendeley bookmarks also correlated strongly with other metrics, including number of domains mentioning titles, OAPEN downloads, and number of Facebook and Twitter mentions. Since this study was conducted with a small sample of titles, future research should conduct the same analyses with a larger dataset.

4.5 Discussion

In this chapter, a more in-depth analysis was conducted on various data related to OA monographs, in particular mentions of the KU pilot collection titles on social media and in citation databases. Twitter and Facebook mentions did not show a significant correlation with citations in citation databases. Although these monographs had more extensive coverage on Twitter (where all 28 titles were mentioned) than Facebook (where only 20 titles were mentioned), only Facebook mentions showed a significant correlation with Twitter mentions ($r=0.52$ with $p < 0.01$). In line with the argument made by Hammarfelt (2014), this may be because the tweets mentioning these titles were advertisements from publishers and booksellers. In fact, most tweets came from the Twitter accounts of KU and the

repository Unglue.it. As such, they are not necessarily reflective of organic interest in these titles among scholars.

TCAT software was useful in this study. Publishers can use this tool to track conversations about their titles and to identify Twitter accounts that are key for disseminating their titles. The tool can be set to capture tweets that mention the titles before their publication. Unfortunately, the KU pilot collection titles were published two years prior to this study, before mentions of these titles had begun to be captured.

These types of data from social networking services such as Twitter and Facebook are valuable for researchers. Most of these platforms provide services for free to their users. However, access to their users' data is usually not free, because most social media platforms view this data as a commodity (Kinsley, 2015). In addition, the limited datasets that these platforms share through their APIs are usually not found to be useful (Alaimo, 2018). Therefore, researchers are typically forced to use commercial third-party services to access complete historical datasets.

No blog posts referencing the KU pilot collection titles were found on the ResearchBlogging portal. Wikipedia articles that referenced 13 of these titles were found, although most of these articles did not direct users to the titles' free PDF downloads.

Among the three book rating platforms (Google Books, Amazon and Goodreads), only the book review site Goodreads provided more than a handful of ratings for the KU pilot collection titles. Goodreads can be a useful proxy for presenting OA monographs, because this site's large number of users generally consult book reviews on this platform before making a purchase. However, across these three review and discovery platforms, only the De Gruyter titles were clearly marked as OA.

The correlation analysis showed a strong positive relationship between Mendeley bookmarks and other metrics, in particular with citations on citation databases. This may be because users who stored monographs in their Mendeley libraries had downloaded them with the intention of citing them later. To understand the relationship between these collected metrics, a classification is needed from the

access of the monograph to its citation. Using this classification, it is possible to shed light on monographs' trajectories across knowledge landscapes (Neylon, 2014).

4.6 Issues

During the analyses conducted for this chapter, two main problems were encountered with monographs' accessibility: incomplete and incorrect monograph records and a lack of awareness of titles' OA status. Since some titles had different ISBNs for each electronic file format, platforms often did not have a complete set of ISBNs for each title. In addition to missing ISBNs in repository metafiles, some titles also had incorrect ISBN entries. Another problem was the titles' publication year. Some titles publication year was indicated as either 2013 or 2014 on different platforms. Sometimes, no date was given at all, which made it even more difficult to identify titles with same name, particularly, for example, where a thesis and a book had the same title. For example, citations for "Composing the Party Line" dating from 2012 and 2013 were excluded because they were citing the monograph author's PhD dissertation, which had the same title, but was published in 2004.

Another issue that was encountered was inconsistent citations of author's names. For example, in some metadata, the author of "Ever Faithful" is written as "Sartorious", and in others it is written as "Sartorius". The same problem was encountered for James Tully, author of "On Global Citizenship": in some references, it is written as "Jim Tully", or sometimes the title is referred to without mentioning the author's name. These problems made it difficult to query these titles.

Another problem was the lack of information on the titles' OA status. For example, Wikipedia articles referred to these titles without providing a link to the free PDF downloads. If Wikipedia provided an interface where articles' authors could check the titles, as they do for the ISBNs, it would play an important role in the dissemination of these monographs to a wider audience. In Google Books and Amazon, apart from the De Gruyter publications, links to the titles' free downloads were not provided.

4.7 Conclusion

This chapter described and examined various interactions with the 28 KU pilot collection titles on social media and in citation index databases. These interactions

included mentions on the social network platforms Facebook and Twitter; citations on blog pages and Wikipedia; reviews and ratings from the book platforms Amazon, Google Books, and Goodreads; bookmarks on the Mendeley platform; annotations on the Hypothes.is and PaperHive platforms; and the citation databases Scopus, WoS and Google Scholar. As a case study, it also explored the dissemination on Twitter of one UCL Press title.

During the analysis of the social network data, more Tweets than Facebook posts that mentioned these titles were found, which is in accordance with the findings of previous studies on journal articles (Zahedi et al, 2014; Costas et al., 2015a; Xia et al., 2016). In addition, Twitter covered all 28 titles, while Facebook covered only 21. More than half of the tweets mentioning these titles were produced in the first four months following the titles' publication. The correlation between social network mentions and other metrics was weak, although Facebook mentions had a stronger correlation with other metrics than Twitter mentions.

The network analysis conducted on the tweets that mentioned the UCL Press title showed how information regarding this title was disseminated in the Twittersphere, which helped to identify the different network clusters discussing the title. Different types of network analyses revealed the Twitter users who were key for the dissemination of the title across different network groups.

As has been found for journal articles, Mendeley data showed the highest correlation with other data sources. Again, similarly to findings for journal articles, these data showed that most bookmarks were made by graduate students (159), followed by undergraduate students (42).

Among the three book-rating and reviewing sites, Goodreads had the most ratings of these titles (49), covering 16 books. These data displayed a strong correlation with citation databases and the Mendeley data, suggesting that Goodreads is an important indicator of monograph usage. The citation databases Google Scholar and WoS showed high coverage of these titles, and also high positive correlation with each other. This indicates that Google Scholar, as a free citation database, is a viable data source in monograph research.

In contrast to the findings of Shuai and colleagues (2012), Wikipedia citations did not show any correlation with citations in formal scholarly publications.

Unfortunately, no research blogs mentioning the KU pilot collection titles were found through the ResearchBlogging site, and no annotations of these titles were found on the Hypothes.is and PaperHive platforms.

This chapter examined the traces of usage of 28 monograph titles that could be captured from social media and citation databases. Since these data sources reflect different types of usage, they need to be categorized under defined type of acts in order to understand better the journey of OA monographs in the digital realm. The correlation analysis showed the extent to which these data sources are related. However, to understand these data sources and their relationships with one another, it is necessary to interpret what they indicate.

The following chapter investigates the reasons behind the relationships between these indicators. It begins with the results of the correlation analyses and uses citation and social theories to interpret these acts of mentioning and use.

5 Interpreting Metrics on Monographs

The previous two chapters reported on the collection of data from different sources and presented the results of correlation analyses conducted on these indicators. The issues encountered during the data collection were discussed. This chapter begins with a critical review of the altmetric categorization of different indicator types and suggests a new categorization for these indicators according to the type of act performed on social media. These acts are then interpreted according to citation theories and social media theories, based on Haustein, Bowman and Costas (2015). In order to adapt Haustein et al.'s study to monographs, the study's design is first discussed, and subsequently the categories of acts are redefined. To date, most altmetrics studies have aimed to identify social media indicators and their relations with one another as they relate to journal articles. Similar approaches have been taken for monographs in the preceding chapters of this thesis. However, precise interpretations of patterns of social media usage of scholarly content, based on user characteristics, social factors or social media platform attributes, are lacking.

Following Haustein and colleagues' steps, the causal chain framework adapted by Ngai, Tao, and Moon (2015) from Mohammed, Ferzandi, and Hamilton's (2010) model was applied to interpret the acts on social media using various social theories. Ngai and colleagues' framework is helpful in investigating factors affecting these acts, since the acts are outcomes of various factors.

This chapter aims to answer the following questions:

- How can we interpret acts relating to OA monographs? And, more specifically:
- How can we uncover the dynamics behind these acts?
- Which social theories are useful for interpreting these acts?

In the following sections, the applicable social theories are examined, and their application in the interpretation of behaviours relating to monographs is discussed.

5.1 Categorisation of data sources according to type of act

Several classifications of the usages of research outputs have been proposed, particularly in the field of altmetrics. Impactstory (2012) classified their data sources into five different categories: recommendation, citation, saving, discussion, and viewing. In addition to these categories, Impactstory also classified interactions according to the actors involved. They divided these acts according to whether they were conducted by academics or member of the public (as shown in Table 5.1). However, it is difficult to distinguish between scholars and the public, because, for example, it is not possible to state that the public holds discussions on Twitter, Facebook and blogs and downloads titles in HTML format, whereas scholars only download titles in PDF format.

Table 5.1: ImpactStory classification of data sources.

	Scholars	Public
Recommended	Citations by editorials, F1000	Press articles
Cited	Citations, full-text mentions	Wikipedia mentions
Saved	CiteULike, Mendeley	Delicious
Discussed	Science blogs, journal comments	Blogs, Twitter, Facebook
Viewed	PDF downloads	HTML downloads

In the article-level metrics (ALM) developed by the PLOS, acts are categorized into viewing, saving, discussion, recommendation and citation (Lin & Fenner, 2013).

Haustein, Bowman and Costas (2015) proposed a framework that classifies acts into three categories: access, appraise, and apply. Their aim was to design a framework that would capture various stages and facets of use and interactions with research objects. Using Bourdieu's (1975) concept of 'agents', these research objects can be scholarly agents, such as researchers, or scholarly documents, such as journal articles or monographs. Haustein et al. also described these acts as having a spiral shape showing increasing levels of engagement with a research object, with 'apply' having the highest level of engagement and 'access' the lowest (Figure 5.1).

Access acts on a research object, which in the case of a monograph would be the document, include viewing the metadata, which may involve viewing the title, the

abstract or a description of the monograph. These acts would also include accessing and storing the monograph, which various platforms and repositories record as download and view counts and is recorded in bookmark counts in reference managers such as Mendeley and Zotero.

The appraise category includes the act of mentioning the document on various platforms, including Twitter, Facebook, e-mail lists (listservs), rating platforms, blogs, Wikipedia, and scientific or policy documents. Lastly, Haustein and colleagues define acts in the apply category as the active use of significant parts of a document, or the adaptation or transformation of the document, which includes applying theories and methods from the document to create new works.

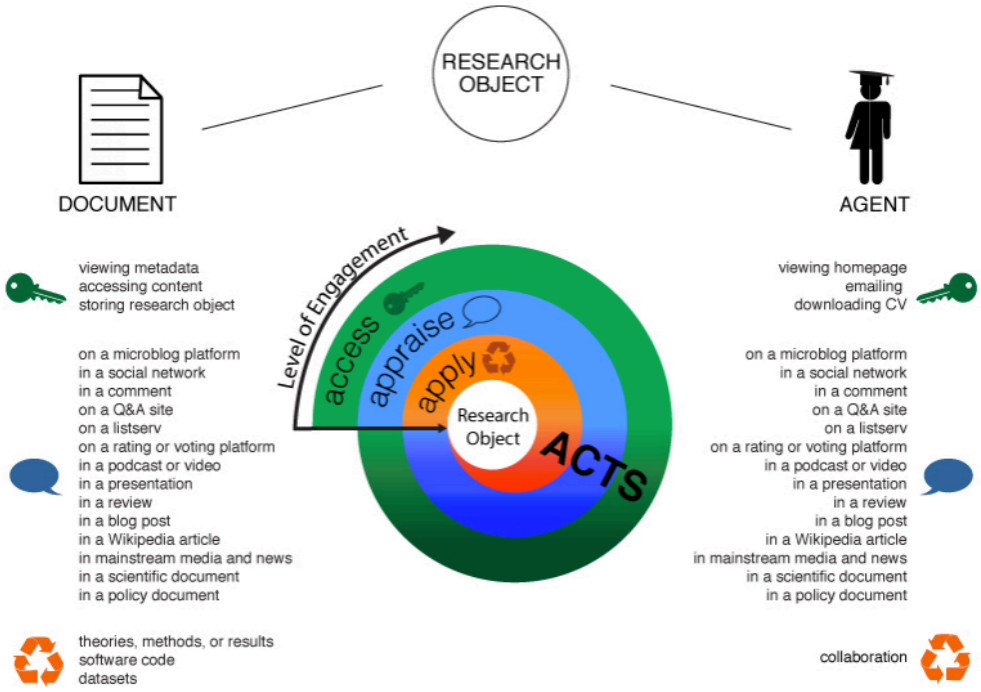

Figure 5.1: Categories and types of acts referring to research objects with their level of engagement. Reprinted from "Interpreting "altmetrics": Viewing acts on social media through the lens of citation and social theories." by S. Haustein, T. D. Bowman, and R. Costas, 2015.

Within Haustein et. al.'s model, the level of engagement increases as one moves through the categories of act from access to appraise to apply, as well as across types of acts within the same category. For example, viewing a document's metadata involves a lower level of engagement than citing it in a Wikipedia article. This is a comparison across categories. Storing an article in a reference manager involves a higher level of engagement than viewing its abstract. This is a comparison within the access category.

An issue with this model is that if level of engagement is defined according to the extent of a user's interaction with a research output, it is difficult to compare acts from different platforms, as these acts represent different things. It is difficult to argue that mentioning a book in a tweet involves a higher level of engagement than reading the document. Most of the tweets observed in this study were for promoting books, which might not have involved reading them, at least in full. In the case of journal articles it seems more likely that reading would precede tweeting as they are shorter. Nonetheless one would hope that reading comes before tweeting in general.

In this study, a different type of categorization is suggested for monographs, which is similar to the models suggested by PLOS and Impact metrics. The categorization's design is based on the indicators used in the previous chapters. However, before introducing these classifications, it is necessary to discuss the types of acts and their relations to one another.

5.1.1 Real acts, defined acts, isolated acts, and associated acts

Previous works on altmetrics and article level metrics have classified each data source under one category of act. The present study approaches this issue by recognizing that each data source may reflect multiple different types of acts. Some types of acts are also redefined here. For example, the fact that a file has been downloaded does not mean that the file has been read. It only means that the file has been downloaded through an act of downloading, which is the act itself that has not been categorized (real act). Since the monograph can be downloaded to be read immediately, or to be stored with the intention of reading it later or not at all, it is not possible to state that downloading a monograph is always a reading or viewing act, although it may sometimes be. Although 'download' is not a defined act, downloading can be classified under different types of defined acts, such as 'access' and 'storage'.

This study proposes to classify data sources that have been used in the previous chapters into six different acts. These are:

- **Access** (Viewing and downloading)
- **Storage** (Downloading and bookmarking)
- **Usage** (Reading, annotating, and bookmarking)

- **Mentions** (Referencing on Wikipedia and mentioning on Twitter, Facebook, blogs, and public web pages)
- **Appraisal** (Reviewing and rating on Amazon, Google Books, and Goodreads)
- **Citations** (Citing in conventional academic literature)

In this classification, mentions and citations in conventional academic literature represent different categories of act. Although citing a research object in academic literature is also an act of mentioning, these two categories are separated to account for citations outside of social media and to enable a comparison between citations from citation databases and citations from social media data sources.

The data sources and their respective types of act are presented in Table 5.2.

Table 5.2: Classification of data sources under defined acts

<table>
<tr><th>Page view</th><th>Download</th><th>Mendeley bookmark</th><th>Annotation</th><th>Wikipedia mention</th><th>Twitter and Facebook mention</th><th>Web pages, blogs</th><th>Book review and rating sites</th><th>Citation databases</th></tr>
<tr><td colspan="2">Access</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td colspan="2">Storage</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Usage</td><td></td><td colspan="2">Usage</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td colspan="3">Mentions</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Appraisal</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Citations</td></tr>
</table>

A data source categorized under one of these defined acts may contain isolated acts, such as a retweeting a friend's tweet that mentions a book. Isolated act does not form part of a series of acts and is probably not a precursor to another act, such as access or storage. In this example to retweet a friend's tweet mentioning a book does not mean that the book is accessed and read, thus will not be cited in formal scholarly publications.

Conversely, some data sources may contain information about different categories of acts. They may be associated with other categories of acts or serve as predictors of other acts. In this study, acts that form part of a series of acts are termed "associated acts". For example, Mendeley bookmarks, which in the previous chapter were found to be the data source most strongly correlated with other data sources, is classified under storage and usage acts. The act of bookmarking carries information regarding associated acts such as access and probably also citation. It is possible to say that

bookmarking in Mendeley is an act within a series of acts, because in order to store and bookmark a monograph, it must first be accessed, and it must also be accessed in order to cite it later. Bookmarking in Mendeley can be classified as a usage act, because Mendeley provides annotation and bookmarking services in addition to its storage service, which means that monographs bookmarked in Mendeley are also used. Therefore, Mendeley bookmark counts can also be interpreted as usage acts.

Another example of an associated act is rating a monograph on Google Books, Amazon or Goodreads. This is classified as an appraisal act. Appraisal acts are associated with usage (reading) and access (downloading or purchasing) acts. In order to perform an appraisal of a monograph, the monograph needs to be accessed and read. Therefore, acts associated with other acts contain information about other acts and display a correlation with these acts.

Returning to the above example of tweeting: all tweets are not necessarily the outcome of an isolated act. Some people who read a book may also tweet about the book, and tweeting then becomes an associated act. However, it is not possible to classify tweeting as an associated act, as is the case for Mendeley usage and monograph rating on Goodreads. This is because Mendeley usage and monograph rating definitely involve at least the access act, in addition to the storage/usage and appraisal acts, respectively.

In summary, this section has recategorized the identified data sources according to six acts. Data sources can fall under one or more categories of act. Each data source as an indicator of act can have associated acts, meaning that it can be an act within a series of acts, which increases its correlation with other data sources. Therefore, it is possible to say that these data sources reveal different layers of acts and may contain traces of different acts.

Apart from containing traces of other acts, the reasons that people perform acts differ both within and across data. For example, interpreting these acts can help us to identify the factors that affect them, making it possible to answer questions such as "What makes people tweet about, bookmark or cite a monograph?"

The following section investigates the dynamics behind these acts and interprets them by probing the intentions and factors associated with them. Even though these data sources are classified within (a) category(ies) of acts, they are interpreted

separately by using citation and social theories, because the motivations for acts can differ according to the data source. For example, a page view of a monograph in HathiTrust can be interpreted differently from an entire download of the same book from OAPEN. These acts' interpretations can change according to the hosting platform, the user, and certain social factors. Following the interpretation of these acts, a causal framework model is proposed to reveal the various factors that cause them.

5.2 Evaluation of acts

Based on Haustein and colleagues' research, and using some of the theories discussed in Ngai et al.'s study (2015), this section describes theories that can be used to understand and interpret these acts that are classified in Table 5.2. Citation theories are discussed and used to understand why authors cite research works in their research outputs or in scientific blogs. Relevant social theories are also used to investigate why individuals mention monographs on social media.

Although the act of citing scientific works in a research output is expected to be linked to Mertonian norms, the reasons underlying this act differ according to the social constructivist theory. On social media, users are not necessarily bound to norms, and many factors affect their actions, including their demographic characteristics, occupation, intentions, perceptions, and social environment, among other factors. Here the perception is related to perceived feeling of a user by using a platform such as ease of use, usefulness, and benefits amongst others. There are also technical features of social media platforms, which determine their choice of platform and actions. Therefore, a combination of sets of theories was used in order to shed light on different aspects of users' acts relating to monographs and to present a more complete picture of these acts. In addition to theories of citation discussed in the second chapter of this dissertation, in the following section social theories, personal behaviour theories, and social behaviour factors will be reviewed.

5.2.1 Social theories

In an effort to understand various acts relating to monographs on social media, social media theories are used in this section. Ngai and colleagues, in their study, gathered a number of theories from 47 selected articles on social media published between 2002 and 2011 (Ngai, Moon, et al., 2015). Although the focus of their

research was on business management and marketing, most of the theories and models covered relate to the socio-psychological behaviour of social media users, and therefore these theories and models were useful in interpreting the motives and acts of social media users in this study. In addition, several theories that are discussed in Ngai et al.'s paper were also used in this study to examine the motives and factors that affected social media users' behaviour in relation to monographs. Based on Ngai and colleagues' work, this section discusses behaviour using two categories of theories: personal behaviour theories and social behaviour theories (Ngai et al., 2015).

5.2.2 Personal behaviour theories

Personal behaviour theories aim to explain the behaviours of human beings at the personal/individual level. Studies using personal behaviour theories focus on user personality, user perception and experience and user intentions. These areas present different aspects of why people behave in the ways they do on the Internet (Amichai-Hamburger, 2002).

5.2.2.1 User personality

A number of scholars have examined users' personalities as they relate to their social media use. They mainly use the five-factor model of personality, which summarizes personality using five factors: openness to experiences, conscientiousness, extraversion, agreeableness and neuroticism (Digman, 1990). Each of these factors is bipolar—for example, the counterpart of extraversion is introversion—and helps to explain various aspects of personality, such as sociability, which in turn contain more specific personality traits (e.g., talkative, outgoing). These studies make it possible to understand users who are engaging in social media activities.

Using the five-factor model, Amichai-Hamburger and colleagues investigated the personality traits of authors of Wikipedia content (Amichai–Hamburger, Lamdan, Madiel, & Hayat, 2008). Other researchers have used other types of personality attributes in addition to these five traits. For example, Zhong, Hardin, and Sun (2011) investigated the associations between social media use and personality attributes in 436 US university students. They found that students who engaged in effortful thinking used social media less often, and those who were more likely to adopt information and communication technologies (ICT) innovations earlier than

others used social media more often. Zhong et al. also found that those who spent more time on social media were more likely to be multitaskers. Additionally, those who spent more time on social media also spent more time on the Internet in general, for study/work purposes or for no specific purpose.

Another factor that affects behaviour is the individual's perception and experience regarding the behaviour. These perceptions and experiences can motivate the individual to do something and continue doing it or can stop them from doing it. A study conducted by Chiu, Wang, Shih, and Fan (2011) investigated individuals' motivations to continue sharing knowledge in open professional virtual communities. They found that enjoyment was critical for the community members' satisfaction and intention to continue sharing.

Users' interactions with technology can also affect their intentions to continue their actions. For example, the success of a platform is directly related to the extent to which the technology it provides fits its users' needs (Ip and Wagner, 2008). However, a user's intentions are not only affected by technology, they are also determined by the user's attitude and the perceived social pressure to engage or not engage in a specific behaviour. Hsu and Lin (2008), in their study investigating users' motivations for participating in blog activity, found that the platform's ease of use and users' enjoyment of blogging were positively related to their attitudes toward blogging. In addition to these factors, social factors, including community identification and attitudes toward blogging, also influenced users' intentions to continue to blog.

Another factor affecting users' intentions is how they identify themselves in specific groups. This factor is relevant because users may act differently in different social contexts according to the social group to which they belong, such as their family, their country of origin, and the sports team they follow. Thus, to improve their self-image, users aim to enhance the status of the group to which they belong. This leads the world to be divided into 'them' and 'us'. This divide can lead individuals to place others in social categories, such as black, white, Muslim, Christian, Jew, student, professional, and so forth. According to Taifel and Turner (1979), following this social categorization, the individual will start to identify themselves with the group to which they belong to. This belonging would provide an important source of pride and self-esteem to the individual, because it provides a sense of belonging to

the social world. Lastly, the individual will perform social comparisons between their group and others' groups. This comparison can lead to competition and hostility between groups leading to an identity. Therefore, users' self-identification with the group to which they belong can motivate them to participate actively in joint activities within the group (Bhattacharya & Sen, 2003). Ely (1994) states that people are motivated to maintain and improve their self-image as a member of a group (Ely, 1994), and this motivation is one of the factors that give rise to the user's intention to use a social networking site (Cheung & Lee, 2010).

Multiple media compete for users' attention, and users select the medium that meets their needs, which may be the desire for information, emotional connection, or status (Cutler and Danowski, 1980; Tan, 1985). Alternative platforms can provide users with a different experience. Chen (2011) found that active Twitter users meet their need to feel connected and form relationships with other people by using Twitter.

During their use of a platform, users seek a fair balance between their input and output (Adams, 1965). Using this idea, Chiu and colleagues (2011) investigated users' motivations to continue sharing knowledge with other members on open social networks. They found that users continue to share knowledge when the satisfaction they obtain from doing so exceeds their expectations.

Although user behaviours such as using a platform, sharing information, and downloading or mentioning a monograph are affected by the user's personality, perceptions, and experience, they are also affected by the user's intentions. A user's intentions to purchase a product, download software, or use specific platforms can be affected by other people's experiences. Online ratings and reviews on social media play an important role in users' choices. According to the Pew Research Centre, 24% of American adults have posted comments or reviews about items they have bought (Jansen, 2010). A 2016 study conducted by the same centre found that half of adults under 50 routinely check online reviews before buying new items (A. Smith & Anderson, 2016). Zhu and Zhang (2010), using data from the video game industry, found that online reviews affect product sales only when consumer reliance on product reviews is sufficiently high. They also found that popular products tend to receive more reviews, which in turn increases their sales. This can be interpreted in relation to the Matthew effect discussed above and in Chapter 2.

5.2.3 Social Behaviour Factors

In addition to personal behaviour factors, interactions among people also play a role in acts observed on social media. In this section, several approaches to social behaviour are discussed in order to understand how users' behaviours are affected by their relationships with the environment of which they form part.

One such social effect occurs when people join a virtual community. Here, their intention in joining the community is not only to gain information or knowledge and solve a problem; they also join to use the platform to meet other people; to seek support, friendship and a sense of belongingness (Y. Zhang & Hiltz, 2003). In other words, they desire to build social relationships with other people inside the community. These social relationships can be seen as a source of means that can be accumulated, in other words, as social capital (Nahapiet & Ghoshal, 2000). In these relationships, reciprocity, trust and cooperation are essential. Chiu and colleagues (2006), in their study of knowledge sharing in virtual communities using social capital, found that social interaction ties, reciprocity, and identification increased individuals' quantity of knowledge sharing.

According to the social exchange theory, an individual's interactions with others are determined by the reward and punishment that they expect to receive from others, which is evaluated consciously or subconsciously based on a cost-benefit analysis model. These rewards can take many forms, including social recognition, money, or even a smile. This theory explains why people help each other, return the benefits they receive, and exchange information and support (Cropanzano & Mitchell, 2005). Using this theory, Blanchard examined how sense of community develops in virtual communities (Blanchard, 2008). Her research, which involved 216 members of five online groups from a list of listservs and usenet newsgroups, showed that the exchange of support positively affects the sense of virtual community.

Because the exchange of support is crucial for the existence and survival of many virtual communities, Blanchard examined group members' support exchange behaviours from the perspective of social exchange theory in order to understand the sense of social virtual community. Here, the sense of social virtual community refers to the participants' feelings of membership, identity, influence, and attachment to one another. Her research, which involved 216 members of five online groups from a list

of listservs and usenet newsgroups, showed that the exchange of support positively affects the sense of virtual community.

5.3 Interpretation of acts on social media

The previous section reviewed several studies that examined the factors affecting acts that occur on social media. This section discusses and interprets indicators belonging to each type of act using the results presented in the previous chapters and the studies reviewed in the previous section (shown in Figure 5.2). The relations between acts will be reviewed in the light of the correlation analysis result obtained in the previous chapter (shown in Figure 5.2).

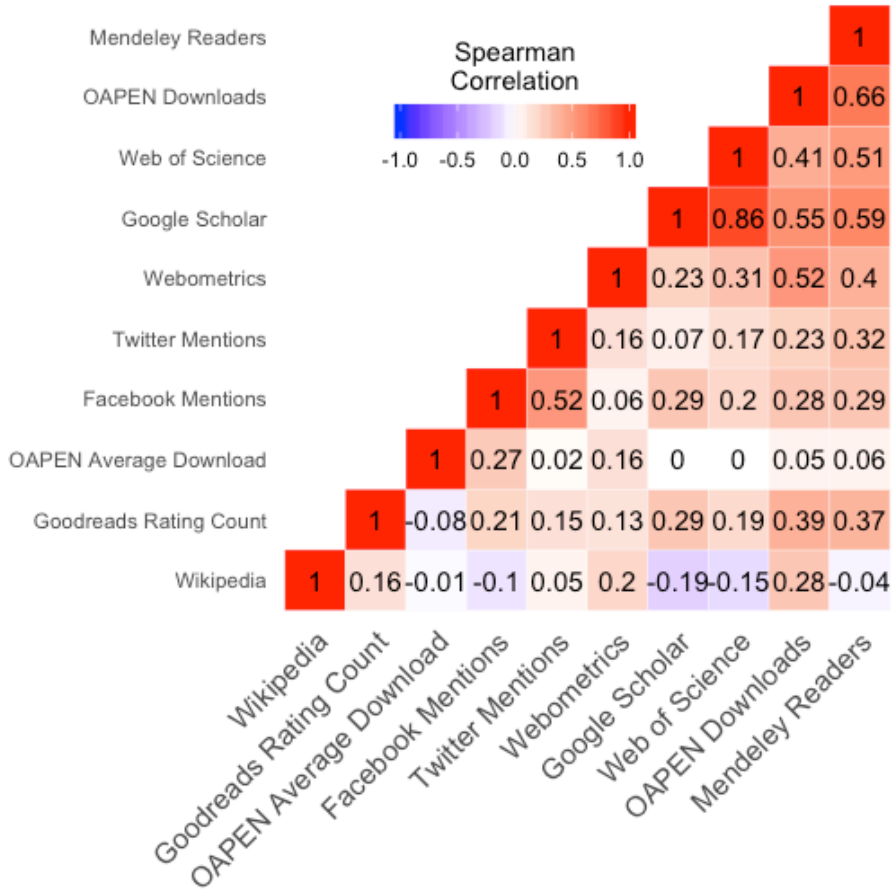

Figure 5.2: Correlation analysis between different data sources.

5.3.1 Access

The 'access' act category includes page views and downloads. As discussed at the beginning of this chapter, the act of access differs from storage and usage. Acts in the 'access' category only involve accessing the content. When a page is viewed, it

is read, but it is not stored. However, when a book is downloaded, it is both accessed and stored, but not necessarily read. Because repositories have different objectives, they present books in different ways and provide their users with different features. Therefore, users' interactions, perceptions and experiences of accessing a book differ across platforms.

Chapter 3 showed that access numbers for the same titles differed across the different repositories. This difference was due to the repositories' differing purposes (i.e., the dissemination vs. the preservation of monographs) and the differences in the discoverability of their titles. However, it was also due to the way in which these monographs were presented and accessed on each platform. For example, JSTOR provides keywords and chapter-level downloads. According to attention economics, discussed in chapter 2, users checking for keywords on JSTOR can save time by using the keywords JSTOR has already extracted for each chapter. The keyword feature also increases the monograph chapter's discoverability on search engines. A study conducted by KU Research on OA monographs on JSTOR for four publishers (UCL Press, Cornel University Press, University of California Press, and University of Michigan Press) revealed that outside of the JSTOR platform, the highest numbers of users came from Google search engine sites (Montgomery, Ozaygen, Pinter & Saunders, 2017). The number of book chapter views was more than 60% of the total number of chapter views and downloads combined. According to attention economics, these results show that the JSTOR platform, instead of downloading PDF contents, also provides contents to users who seek to access them directly via Google's search engine. Although no data were collected on JSTOR usages, users' perceptions, interactions and experiences with a platform's technology are also expected to play a role in their choice of platform for accessing OA monograph content. Therefore, the interpretation of content access can differ across platforms.

Chapter 3 also revealed also that a monograph's language, subject, field, licensing and user location affect its download numbers. This chapter showed that titles were downloaded more from the geographic locations mentioned in their subjects. The six countries with the most downloads of the KU pilot collection titles were those with the most universities in the top 300 globally, according to the Academic Ranking of World Universities (ARWU). Moreover, English-speaking countries seemed to dominate the list of downloading countries. This shows that language is also a factor

affecting the number of downloads. A monograph's license type can also have a moderating effect on access, where OA monographs are accessed more than closed access monographs.

The results in Chapter 3 also showed that a title's subject affects its number of downloads. For example, popular subjects are likely to be downloaded more, as in the case of the "Biological Relatives" and "Constructing Muslims in France" titles. In addition, a monograph's academic field also affects its access. Monographs within narrow academic fields will be accessed less.

In summary, users' intentions, perceptions, and experiences with a platform affect how they access OA monographs. Moreover, the way users access to the contents and interacts with them are defined by the platform and technologies they employ. Users location, language, subject, field and licensing of a title also affect download numbers.

5.3.2 Storage

The act of downloading, in addition to being classified under the access category, is also classified under the storage category, because when a monograph is downloaded, it is stored until it is deleted. However, the act of downloading is not classified under the usage category, because it is not possible to determine whether downloaded contents were read (in this study, reading is classified under the usage category). An analogy is buying a DVD and watching it. These two acts are completely different. Although one does not throw away a DVD they have bought, keeping the DVD does not mean that it has been watched. In the same way, although reading is to some degree associated with the use of Mendeley, research outputs added in Mendeley are not always read. In fact, a study conducted by Mohammadi revealed that only 27% of Mendeley users had read all of the documents they had saved in their libraries (Mohammadi, Thelwall, Haustein, & Larivière, 2015). In the same way that the act of storing on Mendeley occurs before the act of citation, reading is an associated act of Mendeley usage, because users save a monograph to their library in order to use and cite it. This study's findings showed a correlation between Mendeley usage and citations, which aligns with Mohammadi and colleagues' (2015) findings.

As with the act of access, user demographics also affect storage numbers. The previous chapter showed that most of the users who bookmarked KU pilot collection titles in Mendeley were students, as indicated in Figure 5.3. This observation is also

in accordance with Mohammadi and colleagues' findings (Mohammadi et al., 2015). It is likely that younger people use Mendeley more at the stage of their education. Therefore, for this 28-title sample, it can be deduced that age is one of the factors that affects the Mendeley platform's usage rate.

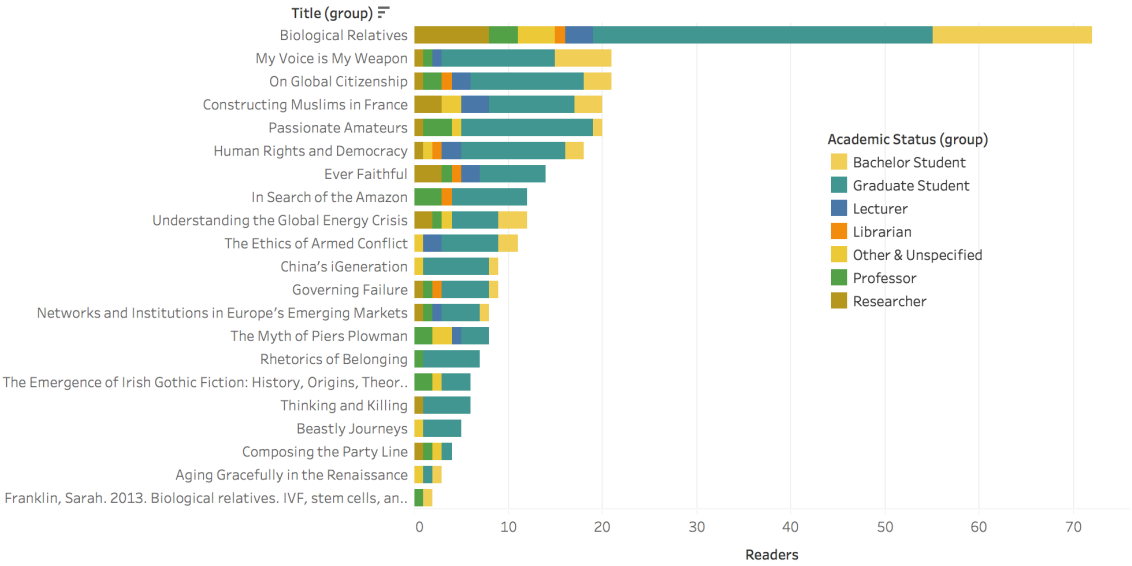

Figure 5.3: Distribution of titles' bookmarks according to users' academic status.

Mendeley offers various features to fit its users' needs and help them perform their tasks. This has made it one of the most used referencing platforms, reaching 2.5 million users in 2013 (Bonasio, 2013). The Matthew effect can be used to interpret the fact that frequently bookmarked monographs attract more bookmarks. Generally, users doing a literature search are in fact searching through other users' libraries. Monographs that are bookmarked more often will be displayed among the search results, resulting in them being bookmarked again, which will also increase their discoverability. Users are likely to add well-known monographs that have more citations or are from well-known authors. This increases these monographs' visibility and the likelihood of them being bookmarked by other Mendeley users.

In addition to Mendeley, the Matthew effect also affects the number of downloads from other platforms. If a monograph has more citations, is from a well-known author, or is published by a prestigious publisher, it will be more discoverable across search engines and in book reviews compared to other published monographs. Thus, these titles are more likely to be visible, leading to them being downloaded more, which would increase their chances of being used and cited in other research outputs. Therefore, these titles have a higher chance of receiving reviews and ratings.

Attention economics can be applied to explain why some users prefer platforms such as Mendeley to organize their documents instead of organizing themselves. Since attention is a scarce commodity, researchers need to reduce the amount of attention expended on searching for documents in their personal or other users' libraries. Users can also run a keyword search across the documents in their personal Mendeley library.

5.3.3 Usage

The usage act category involves acts such as viewing a page, annotating and bookmarking. These acts include page views from the HathiTrust platform; annotations from Hypothes.is and PaperHive; and bookmarks from Mendeley. These acts differ from access and storage acts because they constitute proof that the titles are actually being used. However, it is not possible to identify how much of a research output has been read from these data. Although annotation data make it possible to identify the parts of a research output that are actually read, it cannot be determined whether the parts that are not annotated have been read.

5.3.3.1 Annotation services

In this study, two annotation platforms (Hypothes.is, PaperHive) were investigated. These annotation services display the sections of a document that attract attention. Both platforms offer annotation services in different ways. For example, PaperHive provides annotations for documents that are downloaded to their server, whereas Hypothes.is creates a layer for annotating online web pages, documents, and even offline PDF files.

During this study, no data on annotations of the 28 KU pilot collection titles were found on Hypothes.is and PaperHive. The reason that these services were not used may be that users either do not know about these services or do not perceive any benefit in annotating the monographs on this site. Users may also prefer to annotate their files on their own PDF file readers rather than using these services.

5.3.4 Mentions

The act of mentioning includes referencing books on Wikipedia and mentioning books on Twitter, Facebook, blogs and public web pages.

Wikipedia authoring

Referring on Wikipedia is not classified under the act of citation because Wikipedia articles are not peer-reviewed. Nonetheless, the incentives to write and cite on Wikipedia resemble those that apply in the scientific community.

According to a survey focussing on Wikipedia members' personalities, which involved 139 participants (69 active Wikipedia members and 70 non-Wikipedia members), Wikipedia members showed lower agreeableness, openness, and conscientiousness compared to non-Wikipedia members. Moreover, the survey revealed that introverted women were more likely to be Wikipedia members than extroverted women (Amichai–Hamburger et al., 2008).

Forte and Bruckman (2005) linked the motivations of Wikipedia authors to the sense of credibility that allowed them to assume increasingly central roles as in the scientific community, although there is no direct attribution of authorship. These authors state that although Wikipedia authors receive no apparent credit for their contributions, authors recognize one another and often claim ownership of their articles. For example, the editing history on each article makes it possible to ascertain who created the article and who made the most substantial contribution. Forte and Bruckman also state that authors are incentivized by the possibility of their article appearing on the front page of Wikipedia and the possibility of attracting other contributors to the article.

Authors' acts of contributing to Wikipedia and their intentions to continue contributing are related to their sense of fairness and to whether they feel rewarded for their input. The platform displays author attributions and also displays some articles on the front page of the site. A number of authors display service award badges on their homepages, which is a way for Wikipedia to acknowledge their authors' level of contribution based on the number of edits they have made to articles (Wikipedia:Service awards, n.d.). The platform also indicates each article's access numbers, showing how much attention the article attracts.

In academia, researchers with the same goals and visions build stronger social ties and thus can collaborate on writing articles. However, this is not the case on Wikipedia, due to weak collaborations between authors (Kimmons, 2011). Kimmons states that articles on Wikipedia are not highly collaborative. According to him, the articles do

not reflect much diversity in the construction of their content, but rather represent the work of relatively few people. Kimmons states that the majority of revisions made by users are small or stylistic and therefore may have little impact on the validity of an article's content. A survey conducted in 2007 also found that individuals' motivations for contributing to Wikipedia are mainly related to enjoyment and ideological reasons, such as the desire to make information OA. However, no significant correlation was found between contribution and ideology, and there were no correlations between contribution and social motivations (Nov, 2007).

In this study, eight of the 15 distinct authors with a username that referred to the KU pilot collection titles on Wikipedia had a home page on the platform. Of these eight authors, three had PhDs, one was a university lecturer, and a few were history enthusiasts. It was also observed that the same author referred to the same title in different Wikipedia articles.

Twitter and Facebook mentions

This study found a weak correlation between the number of tweets and the number of Facebook posts with citations, which is in accordance with other studies (Haustein et al., 2014b; Snijder, 2016). Tweet numbers did not have a significant correlation with other data sources either. Tweets from publishers, repositories, and libraries, as well as retweets of these tweets, constituted most of the tweets collected in this study. These tweets and Facebook posts were isolated acts, since organisations who mentioned these 28 titles did not read or cite these titles. They tweeted in order to present these 28 titles.

However, in the long run it is reasonable to expect that with more people reading these books more titles will be shared Facebook and Twitter. If we follow Nederhof's argument (2011) we should wait at least six to eight years after the title is made OA to investigate its world-wide reception. Our findings are also in accordance with Hammarfelt (2014) and Snijder (2016) who found that Twitter was the social media platform with the highest number of mentions, compared to other sources of altmetrics data. Therefore, in the long term, Facebook and Twitter data is expected to be a valuable data source to understand the interactions of people with these titles. These data can also reveal the judgement of readers through sentiment analysis (Pak & Paroubek, 2010).

Mentions on the web (blogs and other web pages)

The correlation matrix in Figure 5.2 shows that mentions on the web (webometrics data) are correlated with downloads, citations, and the number of ratings on Goodreads. Using the Matthew effect, it is possible to state that books with more visibility attract more visibility, although other factors such as demographics may also affect this condition. In chapter 3, a strong linear correlation (r=0.96) was found between the number of mentions from sites based in a particular country and the number of downloads from that country. It was also found that most of the mentions came from university and bookseller sites. Thus, the type and country of the website affects the download rate. It can also be assumed that a book's subject also plays a role in its mentions on the web. Chapter 3 showed that more popular subjects are more likely to be mentioned on websites outside of the academic environment and even on mainstream news sites, such as Jennifer Fredette's "Constructing Muslims in France".

Research blogs

Unlike web pages that mention research outputs but are written by people with different backgrounds, science blogs are mainly written by scientists who write about new research in their own areas of expertise. To provide relevant blog posts for researchers, the aggregator site ResearchBlogging.org was created in 2007. The site aggregates peer-reviewed research posts from many science blogs in seven different languages. The site does not aggregate every post written by its member bloggers, but instead focuses only on those that cite and discuss peer-reviewed research (Zivkovic, 2011). Although no references to the KU pilot collection titles were found on Researchblogging.org, writing and referring in blogs will be discussed with reference to a number of other studies in order to identify factors affecting to these acts on social media.

According to a study conducted by Shema, Bar-Ilan and Thelwall (2012), posts aggregated by Researchblogging.org were primarily from high-impact journals and blogs, including *Science*, *Nature*, *PNAS* and *PLOS One*, indicating the presence of a Matthew effect in research blogging. They also found that the average blogger was male, either a graduate student or had a PhD, and had a Twitter account connected to their blog.

A study involving in-depth interviews with bloggers writing in scholarly contexts from Sweden, the Netherlands and Denmark revealed that these bloggers saw their blogs as a way to disseminate content they wished others to read (Kjellberg, 2010). In addition to disseminating information, they were also motivated by the desire to improve their writing skills and articulate their ideas. Another motivation for bloggers is the feeling of connection. According to Kjellberg, bloggers value the comments on their postings. This feeling of sharing and being connected, which has to do with self-representation and making themselves known, can be explained with reference to their intention to increase their social capital. Several of the bloggers revealed that being surrounded with interesting and interested people is important for the creation of a good research environment. Blogs contribute to this by creating opportunities for connection and participation (Kjellberg, 2010).

According to Yardi, Golder, and Brzozowki, although most blogs are discontinued after some time, people's constant desire for attention is a factor that keeps blogging going (Yardi, Golder, & Brzozowski, 2009). Therefore, readers' interest and their comments affect the writer's motivation to write more.

5.3.5 Appraisal

The appraisal act is an associated act that includes, at a minimum, accessing the appraised item (which, in the case of a book, is either the digital or physical copy). In order for an individual to provide a review or rating of a book, they first need to read the book. In addition to their association with access, book review and rating sites are also influential in increasing access numbers. Platforms such as Amazon, Google Books and Goodreads play an important role in providing information that readers can use to decide whether to read a book. The previous chapter showed that the book appraisal act displays a strong correlation with citations and other data sources.

Chapter 4 mentioned that in a survey conducted by Codex in 2012, 29% of Goodreads users learned about the last book they had bought either from Goodreads or from another book-focused social site. Using the psychological choice model to interpret the result of this survey, it is possible to say that reviews are influential in a user's decision to buy or download a book if the user's reliance on reviews is sufficiently high. In line with Zhu and Zhang's argument (2010), these reviews may have a significant impact on sales or downloads, because books that are downloaded more

tend to receive more reviews, and having a large number of reviews makes a book's overall rating seem more trustworthy (P.-Y. Chen, Wu, & Yoon, 2004). Moreover, as the number of reviews and ratings increases, they tend to be more reflective of the book's quality. Therefore, if the reviews are good and there are a large number of reviews, more downloads are likely, and the larger number of downloads will lead to more reviews, which could be explained by the Matthew effect.

Goodreads is not just a review and rating site; it is also a social networking site where users can follow one another (Thelwall & Kousha, 2017). Thelwall and Kousha found that Goodreads librarians and superusers use most of the features provided by the platform. In addition to seeking information, users also build social relationships with other users on the platform. Therefore, it is possible to say that Goodreads tries to satisfy its users' needs so that they will develop affective feelings toward the platform and the community that may lead them to use more of the platform's products and services. Users on the platform get followers based on their comments on books, which help them to acquire a reputation and feel rewarded for their contributions.

The demographic data of Goodreads users show that most are young, female, come from English-speaking countries, and have a university degree. According to Quantcast, data collected on Goodreads users from 8 May 2018 until 6 June 2018 showed that among the platform's 42 million unique users globally, the United States is the country with the most users, with 17.3 million users. Of the US users, 77% are female, and more than half are younger than 35 (Qantcast, 2018).

5.3.6 Citation

The citation act includes referring in peer-reviewed blog posts and research outputs. As discussed in the previous section, Wikipedia references are not included in this category, because references to research outputs on Wikipedia are provided out of interest by users who are not researchers in the relevant field, and these users make these references without adhering to the Mertonian norms. Research blogs, on the other hand, are usually written by people working in the field in question. However, the references on these blogs are usually not based on Mertonian norms. These authors usually provide one or two references, which are usually based on one article. Therefore, only referring in research outputs is classified under this act. Referring also

has access as an associative act, because referrers need to read a book in order to cite it.

Authors referring to books are expected to follow the Mertonian norms. However, these norms only affirm ideals; as the social constructivist approach states, they do not describe realities. Ziman (2002) argues that researchers' behaviour is also governed by many unspoken rules, which vary across disciplines and countries. Therefore, when referring to a research output, the behaviour of an author is also affected by their demographics, characteristics and social factors. For example, across countries where universities' evaluation and promotion criteria differ, the strength of the Mertonian norms will likely also differ, which will affect how authors refer in their research outputs.

In summary, this section interpreted acts related to OA monographs under six categories in light of findings of this study and those of previous studies. These interpretations have made it possible to identify factors affecting acts. These factors can be categorized as social factors, user attributes, user characteristics, platform attributes, citation theories, and monograph attributes. The next section uses a model approach to understand the dynamics behind acts and how these factors come into play.

5.4 Causal chain model

The causal chain model used in this study is an adapted version of Ngai and colleagues' model, which is derived from the input-moderator-mediator-output model developed by Mohammed, Ferzandi, and Hamilton (2010). Ngai et al.'s framework consists of antecedents (as inputs), moderators, mediators, and outcomes (outputs) and explains the causes and results of user behaviour in the social media realm. The framework is shown in Figure 5.4.

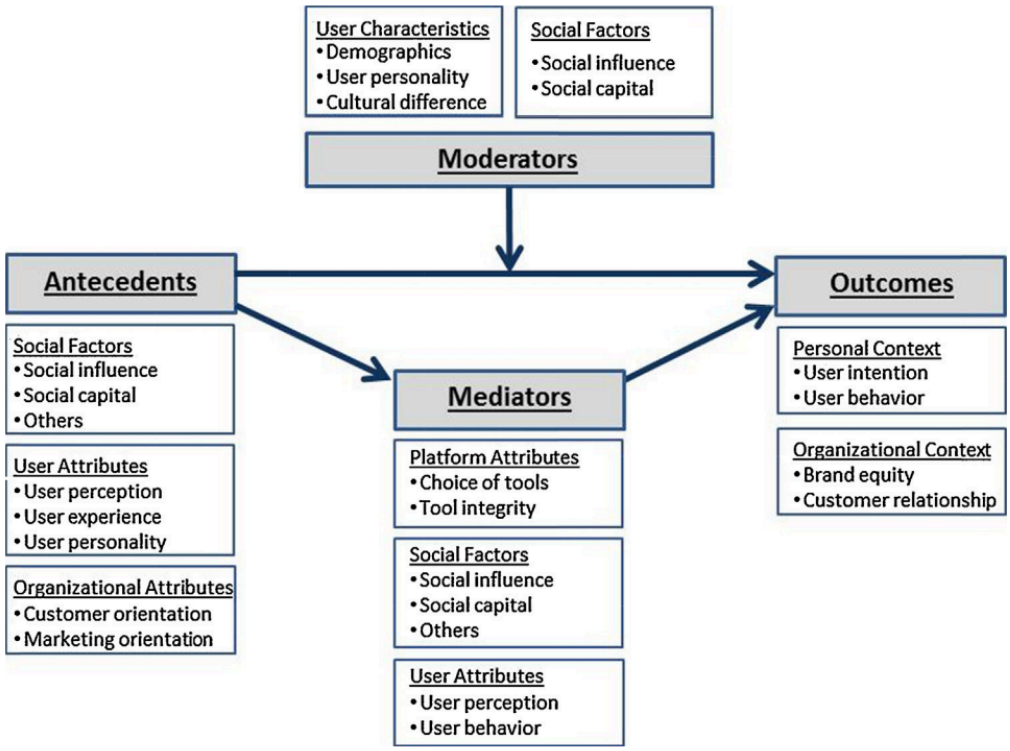

Figure 5.4: Causal chain framework developed by Ngai et al.

In this model, the antecedents are input variables that lead to the outcomes. The mediator factors explain the causalities between antecedents and outcomes, whereas moderators affect the direction and/or strength of these causal relationships. Thus, an antecedent is a stimulus that precedes a behavioural outcome. For example, using this model, it is possible to analyse the relationship between a researcher's opinion about a book after they have read or heard about it and the act of referring to that book. Here, the input variable (antecedent) is the researcher's opinion about the book, and the behavioural outcome is referring to this book. To understand the causality between these two variables, a mediator factor is necessary. In this case, the Mertonian norms are the mediating factor that determine whether an author will refer to the book or not. However, as discussed in relation to the citation act, Mertonian norms are ideals and vary across countries, institutions, and cultures. Thus, demographics and social factors affect the strength of this relationship and therefore they act as the moderator factor. The difference between moderating and mediating factors also explains the relationship between Mertonian norms and social constructivist approaches. Therefore, the relationship between a researcher's opinion about a book and their act of citation can be illustrated as in Figure 5.5 below.

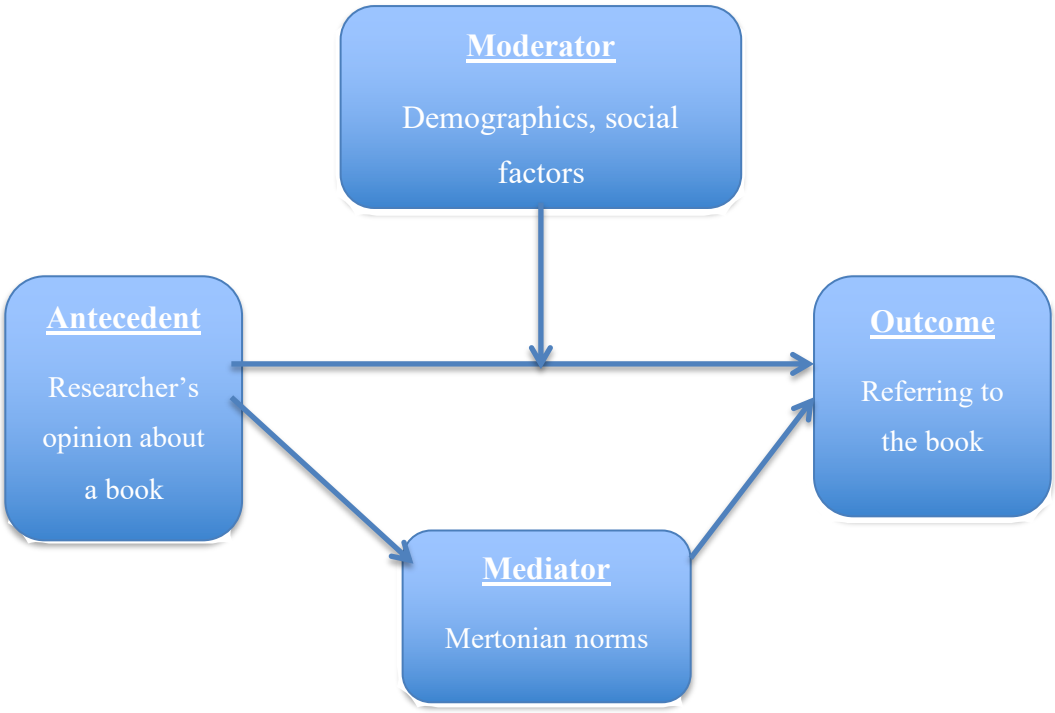

Figure 5.5: Causal chain framework for referring to a book.

In this model, in addition to opinions regarding a research output, antecedents can also be social factors and user attributes. Here, user attributes are social media users' perceptions, experiences or personality. Users' perceptions, as discussed in the interpretation of acts section, are users' feelings regarding the ease of use and usefulness of social media. User experiences comprise a user's involvement with and time spent on a social media platform. For example, the study conducted by Thelwall and Kousha (2017) showed that Goodreads users do not only use the platform as a review and rating site, but also use it to socialize. Therefore, the site fulfils users' need to connect with others by making it possible for them to follow other users. In relation to personality, Amichai-Hamburger and colleagues found that introverted women were more likely to be Wikipedia members than extroverted women (Amichai–Hamburger, Lamdan, Madiel, & Hayat, 2008).

Another example of using this model is to understand the causality between a researcher's opinion about a book that they have read and the act of tweeting about that book. In this example the input variable is again the researcher's opinion about the book, and the behavioural outcome is tweeting about this book. The mediator factors that will determine the user to tweet about a title are the attributes of the Twitter platform, such as user-friendliness, and it being widely used. This tweeting

can be moderated by the personality of the user. Twitter users who are more frequently tweeting and likely to read tweets about intellectual pursuits are found to be higher in openness (Marshall et al., 2018). Another moderating factor can be social capital (Chiu et al., 2006) where the user expects to strengthen his/her social network by tweeting. The relationship of this example can be illustrated as in Figure 5.6.

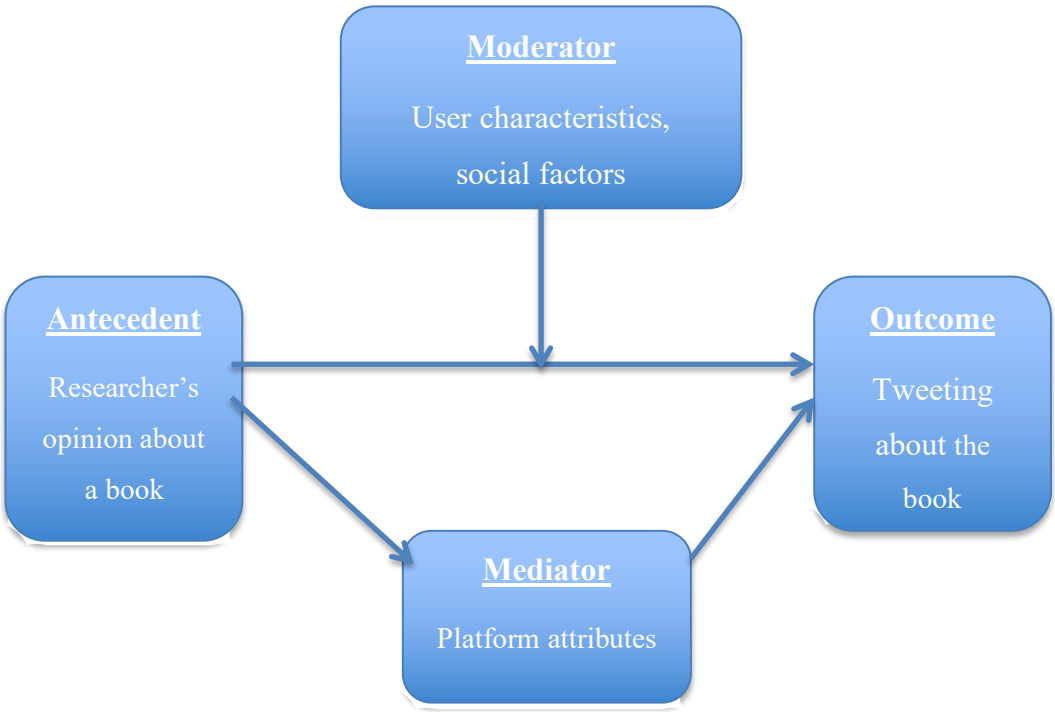

Figure 5.6: Causal chain framework for Tweeting about a book.

Mediators are variables that explain the causal relationships between antecedents and outcomes. Mediators' dimensions can be factors such as the Mertonian norms, platform attributes, social factors, and user attributes. As reviewed earlier, the Mertonian norms are ideals that users are expected to follow when referring to an article. The platform attributes, on the other hand, consist of the selection of tools that a site presents. These factors have the causal effect of antecedents on expected behaviour. Returning to the Goodreads example, the platform offers users the possibility to connect with other people on the platform. However, on platforms other than Goodreads, such as Amazon or Google Books, users may not have a need to connect with others. Thus, the platform as a mediator can change the causal relationships between antecedents and outcomes.

Other mediators can be social factors, such as social capital, social influence and social ties, which are used to explain user intentions and behaviours. Ngai and

colleagues also employed user attributes—comprising users' perceptions and user behaviour—as mediators. They state that although users' perceptions are used exclusively as an antecedent in numerous studies, they are also used as a mediator. For instance, Kwon and Wen (2010) showed how users' perceptions affect users' actual acceptance of social network services. Hossain and de Silva (2009) classified customer attitudes towards the use of social media as a mediator between the inputs of perceived ease of use, perceived usefulness, and influence of ties and the outputs of behavioural intention that led to actual usage. Moreover, in the case of Goodreads, Thelwall and Kousha (2017) found that superusers use most of the features of the site, which increases their engagement with the platform.

In contrast to mediators, moderators are types of research variable, either qualitative (e.g., sex, race, class) or quantitative (e.g., age), which affect the strength of the relationship between antecedent and outcome. The moderators used in social media research can be classified as user characteristics and social factors. User characteristics can be demographic variables, such as gender, age, income, location, or education. Social factors can be social influence or social capital. These user characteristics and social factors are called moderators because they determine the strength and direction of antecedents' influence on expected outcomes.

To summarize the distinction between moderator and mediator, the mediator variable is the middle variable between an antecedent variable (AV) and an outcome variable (OV) (Baron & Kenny, 1986). The purpose of the mediator variable is to explain the relationship between the AV and the OV. The AV does not directly influence the OV; rather, the AV indirectly influences the OV through the mediator variable (as shown in Figure 5.5). For example, social class (AV) positively influences education (mediator variable), and then education positively influences health-screening expenses (OV). When the effect of education is removed, the relationship between social class and health-screening expenses disappears.

The moderator variable, on the other hand, is a third-party variable that modifies the relationship between the AV and the OV. The purpose of the moderator variable is to measure the strength of the relationship between the AV and the OV. For example, if age is a moderator variable between social class (AV) and health-screening expenses (OV), then the relationship between social class and health-

screening expenses may be stronger for older men and weaker for younger men (Fung, 2015).

Finally, the outcomes are the expected results generated by antecedents under the influence of mediators and moderators. They can be in one of two dimensions: personal context and organizational context. Personal context comprises user intention and user behaviour. In Ngai et al.'s framework the organization context relates to brand equity and customer relationship. Therefore, the organization context is out of the scope of this study.

In light of the findings presented in this chapter, this study proposes the model shown in Figure 5.7.

Using this model, it is possible to expose the causal relationship between hearing about or reading a book and mentioning it on Twitter, reading/writing comments about it on Goodreads, bookmarking it on Mendeley, or referring to it in an article.

5.5 Conclusion

The purpose of this chapter was to define categories of acts related to monograph usage and to classify the data sources that have been used in this study into these categories. Using various theories and approaches, each of these acts was investigated and discussed. In light of these discussions and findings, an adapted version of the causal chain model was developed to explain monograph usage. The proposed model is shown in Figure 5.7.

Future studies, including surveys and analyses using a larger sample of titles, are needed to confirm the interpretations presented above. With this chapter, the data collection, analysis and interpretation is concluded. The following chapter review the findings and the implications of this study. Moreover it identifies the challenges in discoverability and identification of OA monographs as well as the issues related to collecting and interpreting data related to their usage.

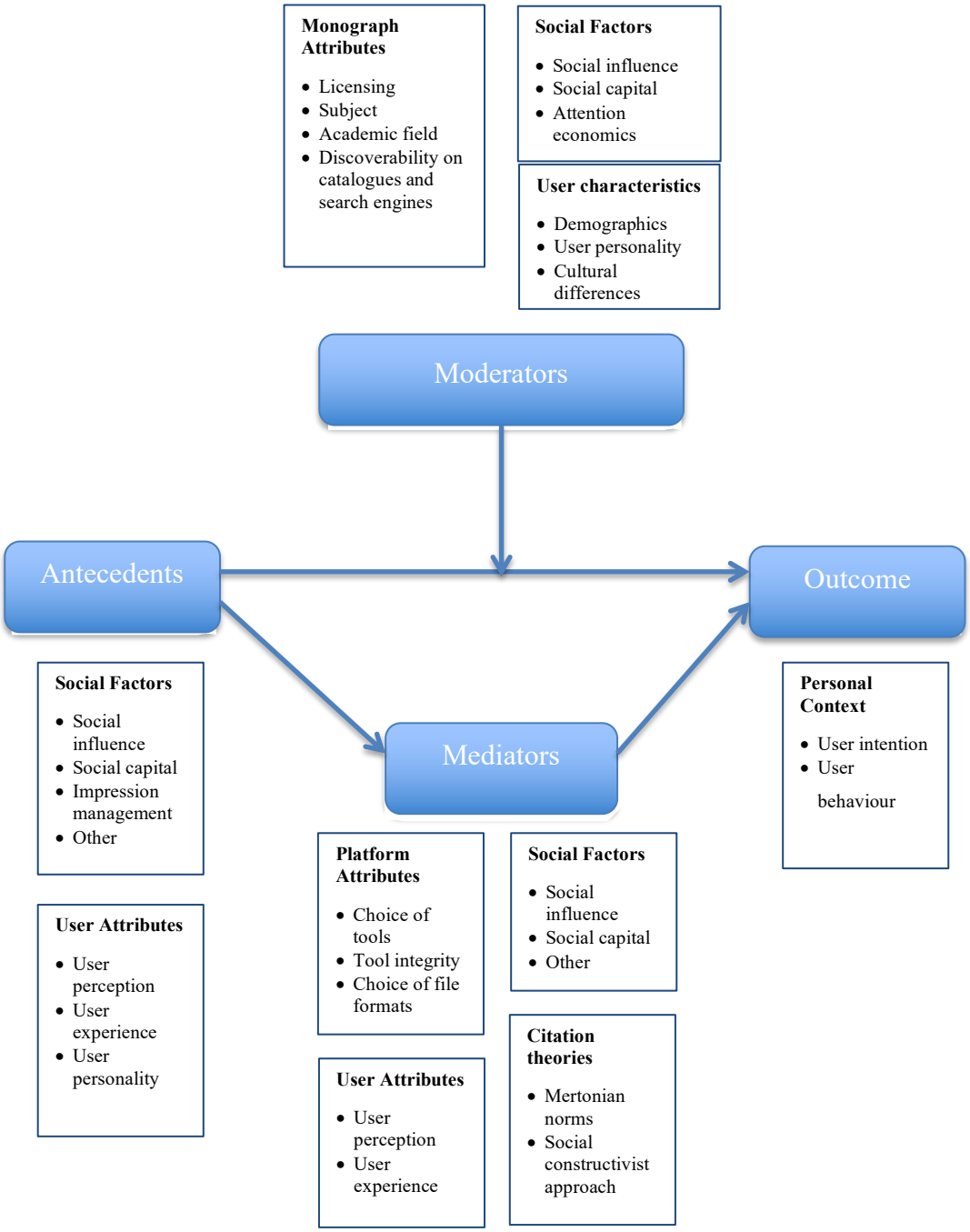

Figure 5.7: Proposed causal-chain framework model.

6 Discussion and Conclusion

This chapter begins by answering the study's research questions using the research findings. It then discusses the issues related to OA monographs, such as discoverability and access issues, identification problems, and challenges in data collection and interpretation. In light of the findings and the issues encountered, the chapter discusses how OA monographs differ from journal articles. Subsequently, it discusses the limitations of this study, the implications of the findings, and suggestions for future research. Lastly, the conclusion summarizes the main findings of this research.

6.1 Summary

The objective of this study was to understand data captured on the Internet related to OA monographs. As part of the study, an extensive range of data related to 28 titles from the KU pilot collection was collected, reflecting different type of acts. Subsequently, these data were analysed, classified and interpreted to provide a detailed overall picture of the dissemination of OA monographs, beginning from the point they were made OA. Lastly, factors affecting interactions related to monographs were identified and a causal chain model was proposed to reveal the dynamics of these interactions. This model can also be applied to other types of research outputs.

6.1.1 Answers to the research questions

To understand the footprints related to OA monographs, the research question "What can rich data reveal about the use of and interactions related to open access monographs?" was posed. To answer this question, three research phases were undertaken, which are discussed in the following sections. The findings of each of these phases complement the findings of the following phase.

6.1.1.1 Research phase 1: Exploring the extent to which use and interactions related to KU OA books can be detected across global digital landscapes.

Seven main categories of data were identified during the study: (1) discoverability of titles using metadata; (2) data on visibility, which are the presences of title names on websites, collected using webometrics methods; (3) access data collected from

reports, including download counts and web traffic statistics; (4) usage data from Mendeley and annotation services; (5) social media mentions, including mentions on social networks, Wikipedia and scholarly blog references; (6) appraisal data collected from book review and rating sites; and (7) citation data from conventional research outputs collected from citation databases.

Chapter 3 discussed the discoverability, visibility and access of OA monographs and how these three aspects were related to one another. In the discoverability section, the metadata of repositories and directory indexes were reviewed. These data are generally collected by platforms such as BASE and OpenAire to improve the discovery of these titles on the Internet and on library catalogues.

The visibility of the titles on the web was examined using a webometrics approach. First, the web presence of each title was assessed by counting the number of web resources in which the title was present. Subsequently, the URL addresses of these identified web resources were examined according to country, organisation and domain. A content analysis was done on a sample of these identified web resources in order to categorize them.

Chapter 3 also identified two types of access data for monographs: download counts and web traffic statistics. The download counts for the OAPEN repository were gathered from the third-party usage statistics service IRUS-UK and web traffic statistics for web pages on OAPEN. The usage data on HathiTrust were collected using the web analytics platform Google Analytics.

These identified data sources made it possible to detect various aspects of monographs relating to their visibility and access. The study found that the most visible titles were those written on subjects of current interest. Since all of the 28 titles are written in English, most of the 3,238 web resources in which these 28 title names were present were from domains in English-speaking countries (2,334, 72.1%). After removing the country-independent TLDs .com, .net and .org and web resources with IP addresses instead of domain names, 1,084 web resources remained. An examination of these web resources' URLs revealed that the 28 titles were most frequently mentioned in sites from the United States (359), followed by the United Kingdom (264) and Germany (100). The content analysis revealed that

these titles were most frequently mentioned on university sites (20.1%), followed by bookseller sites (19.5%) and scam sites (16.4%).

The access analysis revealed that downloads of the titles decreased during the summer months in the Northern Hemisphere. Besides this, no specific downloading patterns were observed after each monograph had been made OA. During the access analysis, unusual access spikes were detected for some titles. The Google Analytics data made it possible to shed light on the causes of these spikes by examining these users' locations and the sources of the titles' web page views. Although the reasons for the download spikes differed in each case, this type of analysis was found to be valuable. Such analyses can also be helpful in assessing the success of title promotions by retrospectively examining the effects of organised events on monographs' access.

The traffic source feature of Google Analytics also revealed that traffic redirection from social network platforms accounted for less than 3% of the total unique page views. In this study, the results of the traffic source analysis revealed that library cataloguing and search systems drive more traffic to the OAPEN web pages than Google Scholar. This result suggests that these titles are still discovered through library catalogues. This shows the importance of titles being included in the DOAB records and being indexed by academic search engines, including BASE and OpenAIRE, which are used by library catalogues.

Moreover, a positive correlation was found between a country's number of downloads from the OAPEN repository and the number of websites belonging to these countries in which the KU pilot collection titles were present.

Chapter 3 compared the discoverability, visibility and access of the 28 titles between repositories. The repository comparison showed that titles registered on the DOAB had more visibility and access. It was not possible to draw a general conclusion regarding whether visibility was affecting access or access was affecting visibility. However, in the case study that was conducted, a title's visibility in mainstream media was found to have a positive effect on its access.

Chapter 4 investigated mentions on social media. These mentions included data from social networks, including Facebook and Twitter, which were collected using these platforms' APIs; references from Wikipedia, which were collected by querying

on the Google search engine; book ratings and reviews on websites, including Amazon, Goodreads, and Google Books; bookmark data from the Mendeley reference manager, collected using the platform's API; annotations on Hypothes.is and PaperHive; scholarly blog references queried using the ResearchBlogging platform; and citations from citation databases, including Scopus, WoS and Google Scholar.

The findings showed that, for the period between 1 January 2014 and 1 July 2017, all the titles were mentioned on Twitter across a total of 493 tweets, whereas 20 tiles were mentioned on Facebook across 96 Facebook posts. Half of the 493 tweets observed were produced in the first four months after the relevant title had been made OA. On Wikipedia, 23 articles in four languages referred to only 13 titles. Among book rating and reviewing platforms, Goodreads covered 16 titles with 49 ratings, Amazon covered three titles with four ratings, and Google Books provided only one review of a single title.

The Mendeley data covered 20 titles, which were bookmarked by 288 readers. These readers were mostly graduate students (159, 55.2%), followed by undergraduate students (42, 14.6%).

Among the citation databases, WoS covered 27 titles, while Scopus covered only eight titles. However, the free alternative Google Scholar covered all the titles, which suggests that it may be useful for citation analyses of OA monographs. No data related to the 28 KU pilot collection titles were collected from the ResearchBlogging site, which may be because the blogs covered by this site are primarily focused on scientific, technical, and medical research. No relevant data were collected from the annotation services either, possibly because these services are not known to users reading HSS monographs or because users prefer to use their own PDF file readers rather than these services.

6.1.1.2 Research phase 2: Identifying and investigating the relationships between these interactions

A correlation analysis was conducted between the data sources identified in the previous section. A weak correlation was observed between social networks data and data from other sources, although Facebook had stronger correlations with other sources than Twitter did. Among the data sources, Mendeley showed the strongest

correlation with other sources, as has been found to be the case in studies of journal articles. The Goodreads platform, with its book ratings, was found to be a good source of data for analyses of interactions related to monographs, since it displayed a strong correlation with the citation databases and Mendeley data. The citation databases WoS and its free alternative Google Scholar showed high coverage of the KU pilot collection titles. No correlation was found between the data from Wikipedia and other data sources.

To understand these interactions, Chapter 5 classified the abovementioned data sources under six categories of acts, which are 'access', for number of page views and downloads; 'storage', for downloads and bookmarks on Mendeley; 'usage', for page views, Mendeley bookmarks and annotations; 'mentions', for references on Wikipedia and mentions on Twitter, Facebook and blogs; 'appraisal', for book reviews and ratings; and 'citation', for referring in conventional research outputs such as journal articles and monographs. This classification differs from other classifications. In addition, some data sources could be classified into more than one act. This is the case for bookmarking on Mendeley, which is classified as a 'storage' and a 'usage' act.

The classification and correlation analysis helped to understand the association of some acts, such as Mendeley bookmarking, with other acts. Bookmarking in Mendeley is one act within a series of acts. To bookmark a monograph on Mendeley, the title first has to be accessed and stored. In addition, individuals bookmark monographs in Mendeley in order to refer to them later in academic outputs. Since bookmarking in Mendeley is associated with other acts, this act displayed a strong correlation with other acts, including access, storage, and citation. In other studies, although similar correlations were found and some suggestions were made regarding why there was a correlation between two data sources, there was no mention of data sources' associative nature.

In Chapter 4, mentions on social networks, including Twitter and Facebook, showed weak correlations with other metrics such as citations and downloads. This is because, unlike journal articles, where researchers can read a newly published article immediately and subsequently mention it in a tweet, books are generally mentioned in tweets by publishers, repositories and book distributors rather than researchers. Thus, it is not possible to say that social network mentions are a predictor of

citations or an indicator of a monograph's impact (Snijder, 2016), as in the case of journal articles (Eysenbach, 2011).

This study also showed that not all books have an equal chance of success, not because of their quality, relevance amongst other attributes, but also because of how well a publisher launches it, including getting it into all the channels and platforms whether OA or not.

6.1.1.3 Research phase 3: Interpreting the detected interactions using social theories and citation theories and attempting to uncover the factors affecting them.

Chapter 5 interpreted the results obtained from the data sources using citation theories and social theories. Each of the data sources was evaluated and discussed in light of these theories and the correlation analysis results in Chapter 4. For example, bookmarking in Mendeley and referring in Wikipedia were discussed in relation to citation theories. Users' choices between different types and formats of monographs, as well as different types of platforms, were also interpreted in light of social theories and user demographics.

Chapter 5 also discussed how user perceptions and experiences, the platform they use, and social factors determine users' actions and how user characteristics such as demographics, personality and culture affect the strength of this relationship. Subsequently, these factors were grouped into citation norms, user attributes, platform attributes, social factors, and user characteristics, and how they determine user acts was discussed. Some of these factors were observed to cause certain acts, while access, mention or appraisal rate were affecting the strength of this causal relationship.

Finally, these factors were placed in a causal chain model, which was adapted from the work of Ngai and colleagues. This model made it possible to uncover the dynamics behind acts related to OA monographs. Although this model was proposed for OA monographs, it can also be used for other types of research outputs.

6.2 Discussion on the issues related to OA monographs

Another objective of this study was to flag the issues and challenges encountered in the collection of data relating to OA monographs. This section discusses each of these issues, which are classified under three subsections.

6.2.1 Discoverability and access issues relating to OA monographs

Making monographs OA and uploading them to an online repository is not sufficient for them to be accessed; they need to be discovered as well. Chapters 3 and 4 observed that monographs are discovered through search engines such as BASE and OpenAIRE and library catalogues. These intermediaries operate by collecting metadata from OA repositories or directory indexes such as the DOAB. Unfortunately, most of the titles hosted in European repositories are still not registered in the DOAB (Neylon et al., 2018), which makes them less discoverable (McCollough, 2017). Chapter 3 found that some titles were classified under more general subjects, while some were classified under specific subject fields, which makes them less discoverable.

Another issue is that these OA monographs, or information in general, is seen as a commodity on the Internet. In Chapter 3, the content analysis of the 28 KU pilot collection titles revealed that 16% of the web resources in which a title's name and author were present were scam sites. Since monographs are seen as a commodity, scam sites use them as bait to attract users and subsequently collect their email addresses and credit card details. Related to this commodity issue is the fact that most of the monographs' publishers do not provide links to the free electronic versions on their book presentation pages. Some publishers even sell the PDF versions of their titles from their web pages. This approach prevents users from finding out that these titles are in fact OA. One way to overcome this commodity issue and increase the discoverability of OA monographs is to make search engines index the DOAB records, so that when queried, they can display the OA monograph at the top of their results page and make users aware that the monograph is freely accessible.

Chapters 3 and 4 found that most of the mentions on web pages, including references in Wikipedia articles, did not provide a link to the free versions of these titles. This was probably due to the articles' authors not knowing that these titles had

been made OA. Although Wikipedia offers links to the monographs through Google Books, Open Library, or Amazon.com, it was not possible to access the OA content of these titles through these platforms. An inclusion of the DOAB index in Wikipedia ISBN search would provide links to the OA versions of these titles.

Another obstacle in accessing the OA versions of the titles is that on the book review site Goodreads, it is not possible to determine whether a book has been made OA. For example, UCL Press, which promotes OA publishing, has written in the discussion sections of the Goodreads book pages that the titles are actually OA and can be reached via their sites. However, most publishers do not write these kinds of notices for their titles. Unfortunately, Goodreads, which is now part of Amazon, directs users who want to access the content to other online Amazon stores such as Abe Books and Book Depository.

During this study, none of these titles were accessible for free through Amazon.com. They only provide free access to classic books. Among these publishers, UCL Press began to publish their OA monographs on Google Play, so readers can obtain their books via this platform for free.

All these obstacles prevent users, particularly those outside of academia, from discovering that these titles are in fact freely accessible.

6.2.2 Issues regarding the identification of OA monographs

In 2017, Crossref (a not-for-profit official DOI Registration Agency), in cooperation with DataCite (a not-for-profit organisation aiming to improve data citation) began collecting usage data from a number of data source, such as Hypothes.is, blog posts, Reddit, Twitter, Wikipedia and Wordpress.com. Their aim is to provide raw data for analysis for each DOI, without providing any metrics, totals or interpretations.

However, unlike journal articles, one of the problems in tracking monographs is that monographs do not reside in just one repository, and each title from each repository has a different DOI. This is why it is more difficult to track a book's mentions and usage on the Internet. In addition to being hosted on different repositories, titles are also available in various formats. For example, for each different format (PDF, ePUB, MOBI, etc.) of the same book hosted on the same platform, a different DOI is assigned. Some platforms even assign a separate DOI to each chapter of a

monograph. Moreover, in the case of a new edition of a book, these repositories renew all of their DOI records for the book. Thus, in contrast to journal articles, it is possible for a monograph to have tens of different DOIs. There may also be different ISBNs for each format of a title. Therefore, since each title does not have a single unique identifier, it is difficult to track mentions of monographs, especially on social media. However, some publishers use a unique ISBN identifier that encompasses all the different electronic formats of a book.

To overcome this problem, one approach could be to gather and record all the different DOIs under a particular book's ISBN. These records could be hosted on a central organization's server, such as the DOAB or WorldCat, where metafiles from publishers and repositories could be matched. These records could also serve to verify the authenticity of these OA books, and they could provide different format options on different repositories for library catalogues and discovery tools. These records would also be useful in tracking and gathering all the mentions and usages of a title hosted on many repositories. In this way, all the interactions related to a specific title could be tracked on the digital landscape and analysed comprehensively.

Fingerprint of a monograph file

Another issue is to ensure that a monograph tracked and accessed on the Internet is the original version of this monograph. To make sure that monograph content on the Internet is not changed, a monograph file fingerprint can be used. These fingerprints are strings that uniquely identify a file, just like human fingerprints. In fact, the annotation platform Hypothes.is uses PDF fingerprints to identify the uniqueness of PDF files (ISO, 2008). In this way, users can use the annotation platform offline as well.

6.2.3 Issues regarding data collection from different sources

There are various initiatives for collecting data on monographs. One of these initiatives is the Bookmetrix platform, developed by Springer and Altmetric.com. This platform collects usages and mentions on social media for Springer Nature's various ebook collections. Another altmetrics platform is PlumX, which was acquired by Elsevier in 2017, and tracks 4.1 million books and book chapters. In addition to metrics collected from Elsevier's products (Scopus and Mendeley),

PlumX also collects data from different sources, including WorldCat, Amazon.com, Goodreads, and EBSCO.

Although they provide valuable data, both of these services are subscription-based and do not include usages from repositories other than their own products. Since these metrics services are subscription-based, various parties in the OA monograph sector are trying to overcome this aggregation of usage data from multiple source issue. In an effort to integrate OA monographs into the open science ecosystem in a systematic and coordinated way, the HIRMEOS project was developed. The project involves five publishing platforms, and one of its work packages is the metrics service, which is planned to provide altmetrics and citation metrics, as well as a widget to display metrics on partners' websites (HIRMEOS, 2017).

However, there are various challenges in collecting data from these different sources. The main problem is that monograph titles may differ across different repositories or social media platforms. Some repositories may include subtitles or some characters may be changed; for example, a colon may become a dash, which makes tracking less accurate.

Another challenge relates to the ISBN identifiers belonging to a title. Repositories sometimes use different format identifiers. For example, a repository A may use the PDF ISBN, while another repository B may use the print ISBN for the same title. Occasionally, records contain incorrect DOI and ISBN identifiers. Some repositories contained duplicate records or had assigned the same DOI or ISBN identifier to different titles, among other problems. These issues make it more difficult to gather usages for the same title.

As suggested in the previous section, in order to overcome these challenges, an organization such as the DOAB or WorldCat could build and provide a central recording system, which would hold records including monographs' titles, ISBNs, DOIs, formats and a checksum of the file, which is a long string that describe the content of the file and act as the fingerprint of the file. Before building this central recording system, all repositories' metafiles should be cleared of incorrect or duplicate entries. Subsequently, when producing the central records, titles should be recorded from these repository metafiles. To group records under the same title, first ISBN and later DOI identifiers should be used. Subsequently, on the record, titles

with missing, varying, or incorrect ISBN identifier entries should be corrected by using metafiles from different repositories. To match the same title recorded with a subtitle or different characters in other repository metafiles, fuzzy string matching technique such as the Levenshtein distance algorithm[3] can be used.

After matching the titles from different metafiles, the last step would be to fill in the missing attributes of the monograph on the central recording system using the WorldCat database, including identifiers, subjects, and classifications, amongst other attributes. These would help clean, fix and standardize these attributes on the central recording system.

This study found that monographs were mainly mentioned on social media using the title and author's name. This is in contrast to journal articles, where the DOI is typically mentioned. For this reason, it is more relevant to examine mentions using title name and author name, as was done in the third and fourth chapters when collecting mentions and references on Twitter, Facebook, Wikipedia and other web pages.

Using the proposed central recording system, it would also be possible to gather usage data from different repositories. However, there will be challenges in interpreting and displaying access data. The first problem would relate to comparing different accesses, since most repositories track access differently. As explained in the third chapter, to overcome these access challenges, the COUNTER report was developed in order to establish a standard for access statistics. However, unfortunately, not all repositories use COUNTER-compliant reports.

Although combining COUNTER-compliant usage with other usages can provide an overall indication of usages, these types of usages are actually completely different. This method is particularly inappropriate when comparing chapter downloads with whole-book downloads or page views with downloads.

---

[3] The Levenshtein distance algorithm is a measure of the similarity between two strings. In this algorithm, the operations are the removal, insertion, or substitution of a character in the string. For example, the distance between "kitten" and "sitting" is three. The steps involved in transforming 'kitten' to 'sitting' are:

> kitten enshtein distance algorithm is a measu
> sitten enshtein distance algorithm is a measu
> sittin enshtein distance algorithm is a measure

The question that arises, then, is how these metrics can be combined. As discussed in Chapter 5, downloads and page view metrics reflect different things. However, I suggest that they can be combined under the 'access' category. In this way, it becomes possible to combine the number of chapter downloads and whole-book downloads. For example, the interpretation of download counts changes when chapter downloads are compared with whole-book downloads outside of the 'access' context. In this case, dividing the number of chapter downloads by the number of chapters in the book would not make any sense either. It may be that a reader only wants to read the chapter, and not the entire book. Should we interpret a chapter download as the reader having downloaded one fifth of the book? Or how can we compare repository usages which are tracked using COUNTER-compliant methods and usages tracked with other methods? Normally, COUNTER-compliant usages should be lower than web log analysis findings, since COUNTER removes bot usages, consecutive downloads from the same IP address, and other types of usage in order to prevent gaming (Project COUNTER, 2016). Moreover, information is not available on how usage statistics are compiled in the traffic reports of some platforms, such as Google Books. Therefore, it is currently not possible to combine different download usages in an accurate and standard way.

6.2.4 Issues regarding the interpretation of metrics

Different data sources can be combined to generate scores, for example altmetric attention scores for research outputs, or ResearchGate scores for researchers, amongst others. According to Altmetric.com, the altmetric attention score is based on the number of posts that mention an output and the quality of the posts' sources. The company states that they measure public attention, not quality (Altmetric.com, 2017), and these attention scores are calculated according to weighted counts (Altmetric.com, 2018).

Table 6.1 shows the weights of different data sources in their scorings.

Table 6.1: Weight of different metrics used for scoring by altmetric.com.

News	8
Blogs	5
Twitter	1
Facebook	0.25
Sina Weibo	1
Wikipedia	3
Policy Documents (per source)	3
Q&A	0.25
F1000/Publons/Pubpeer	1
YouTube	0.25
Reddit/Pinterest	0.25
LinkedIn	0.5
Open Syllabus	1
Google+	1
Patents	3

According to Almetric.com, in terms of attention, one tweet is worth four Facebook posts, one fifth of a blog article, and one eighth of a news article. However, it is not clear how they devised these weights and how, for example, a blog article can be seen as worth five times more than a tweet. For example, an article may be posted on a blog that is followed by only five people, and a single tweet may be posted by a user with 100,000 followers. In the case of a book, the title may be discussed on social media without the use of a DOI, as is usually the case, or only one chapter of the book may be discussed, without the book's title being mentioned. The best way to present metrics about a research output would be not scoring it but providing the metrics as they are, because scores could otherwise become targets to attain, which could have a harmful effect on scholarly communication and academia. As indicated by Goodhart's law (Strathern, 1997):

When a measure becomes a target, it ceases to be a good measure.

6.3 Differences between OA monographs and journal articles in terms of data analysis

A special subsection is required to emphasize the difference between monographs and journal articles in relation to data analysis. As the introductory chapter mentioned, monographs differ from journal articles in many ways. There is more

diversity and more scarcity in the monograph market (Adema, 2010), which results in different business models being used, books being presented in different ways and in different formats, and books being shared via different repositories. Although these differences make the interpretation of data related to OA monographs difficult, the diversity and scarcity in the monograph market may in fact lead to the development of new ideas, such as new ways of publishing, hosting and analysing metrics for OA monographs.

As discussed in the previous section, this study found that unlike journal articles, which are usually identified by just one string after publication, an OA monograph can be identified in many ways depending on its format, the repositories in which it is hosted and how it is shared (as chapters or as a whole book). These issues pose unique challenges in tracking activities relating to OA monographs.

In addition, this study also showed that social media users do not share monographs' DOIs or ISBN identifiers like they do with journal articles, which makes monographs more difficult to track. They rather choose to share the title and the author's name. The unique situation surrounding monographs should be taken into consideration when gathering data about interactions related to OA monographs.

On the other hand, it has been observed that monographs' usage patterns are also different from those of journal articles. Because it takes longer to finish reading a book and books can be used and remain valid for a longer time, their mentioning patterns differ from those of journal articles. According to Nederhof (2011), to track and analyse citation data for monographs, a period of six to eight years is suitable. Mentions of a journal article on Twitter usually appear in a short period of time after they are made OA. However, it is not possible to predict the number of citations that OA monographs will receive in the same way.

In conclusion, capturing and analysing data related to OA monographs using a single model is not feasible at present. New methods of organizing the available data in a systematic and generalizable way are needed. Therefore, new models that aim to provide customized analytics for publishers are expected to emerge.

6.4 Limitations of this study

The main limitation of this study is its generalizability. Although they were pointed out during the study's design, three limitations are discussed here: sample size, platform specifics, and interviews/online surveys with readers.

The sample used in this study consisted of 28 titles in the fields of anthropology, history, literature, media and communications, and politics, provided by 13 scholarly publishers. These titles were published in 2013 and 2014, which is quite recent in relation to the usage of books. Therefore, the results of this study cannot provide a comprehensive indication of how metrics have changed and affected one another over longer periods. In addition, the titles used in this study were all published in English. The number of publishers and the number of subject fields was limited. Therefore, it was not possible to compare and comment on how language, publisher size and field of study affected the access to and visibility and distribution of OA titles. A comparison with titles behind a paywall would also help to determine OA's usefulness in disseminating knowledge. Another issue related to the use of a small sample size is the lack of data on annotation services. It could also have been better if the Twitter data links could be captured from the beginning of the titles publication. This way we could have an understanding on how these 28 titles were spread across Twittersphere.

The study's second limitation was the number of platforms investigated. Although this study investigated the most popular platforms and covered most of the data, there may be other platforms that offer similar services. A comparison between platforms with different features which cause different user behaviours and outcomes is needed. This kind of analysis would also help to develop more granular results by using the causal chain model in order to highlight the features needed by users.

Finally, the third limitation was the lack of interviews and online surveys with readers, which made it impossible to triangulate the interpretation of different data sources using the causal chain model. Conducting an online survey would also provide information on where users were coming from, where they learned about the titles, what their intentions were in using the platform, and whether there were any other associative acts they were planning to perform.

6.5 Implications

The results and analysis presented in this study have various implications related to the discoverability and visibility of and interactions related to OA monographs. Another important implication of this study relates to the interpretations of data sources relevant to OA monographs.

This study fills a gap by providing a set of methodologies and approaches for assessing the discoverability, visibility and access of OA monographs. In addition, the interpretation of data sources related to OA monographs not only filled a gap in understanding interactions related to OA monographs but also furthered the research begun by Haustein and colleagues (2015) on the interpretation of altmetrics data.

The causal chain framework has created a new path for future studies. It should provide a useful method to explain acts related to research outputs and to uncover the dynamics behind these acts, which will provide information on why and how these titles are accessed, used, shared, appraised and mentioned.

In addition to the theoretical and research implications, this study has made practical contributions.

By using different sources of access data, publishers will be able to gain insight into the reasons behind abnormal access spikes for titles and measure the success of their book promotions. Likewise, libraries will benefit from the study, as they can determine how these titles are accessed within their institutions, which will be useful in their decisions regarding which titles to support.

This study aimed to benefit small and independent publishers in addition to established ones. To do so, it has provided publishers with the methods and approaches necessary to collect visibility, access, usage and mentions data related to their titles. It has also indicated how to analyse these in order to gain insight into the relative performance of individual books and collections, by benchmarking them against other titles, mapping their uses over time, and displaying their performance. It is hoped that these methods and approaches will help publishers to increase the discoverability and visibility of their titles. Even if they lack the necessary technical abilities, this study is expected to help publishers to determine the extent of the information they can request from analytics organisations in this area.

Because of the differences between journal articles and OA monographs identified in this study, searches for interactions related to OA monographs should be performed differently depending on the interested party. Established publishers could use the methodologies in this study and develop their in-house analysis to gather information about their own titles. Other publishers with insufficient technical skills could benefit from this study by identifying the extent of the information they can obtain and expect these methods and approaches to be employed by third-party analytical service providers. In this way, it is expected that this study will also contribute to the emergence of small individual organisations working on OA monograph analytics, due to the tailored services they could provide to individual publishers.

By tracking and examining their users' behaviours using the causal chain framework, platforms will be able to develop new tools and features to facilitate their users' interactions.

Finally, as a whole, this study has aimed to provide the necessary methods and approaches for stakeholders in the OA monograph domain, including publishers, repositories, funders, libraries, platforms and policy makers, to examine users' behaviour and understand their needs in order to make informed decisions to increase the effectiveness of scholarly communication around OA monographs.

Open access monographs are important in spreading knowledge globally. To properly disseminate OA monographs, it is necessary to gather and interpret data related to these titles precisely. As outlined in this study, there are specific challenges related to their tracking. Using a more standardized approach to identify monographs is necessary to track them in a more comprehensive and efficient way. Hopefully, this study's identifications of these challenges will be the start of a search for this standard approach.

6.6 Suggestions for future research

Considering the limitations of this study, this section suggests that future studies include a larger sample of OA monographs as well as monographs behind paywalls. Studies examining the global usage of OA monographs in relation to how OA benefits certain regions would be of interest in the future.

Studies with a larger sample size will also make it possible to compare how the dissemination patterns of these monographs change over time. Studies that include titles published in different languages, on different subjects, and at different times by a number of publishers would indicate how these factors affect their visibility, distribution and access.

By applying social theories, future studies conducted on different social media platforms with similar services would reveal the dynamics behind users' interactions with the platforms. Interviews and online surveys are also needed to triangulate data related to these usages and to evaluate findings obtained from social theory applications. These studies would help to develop and further refine the causal chain model proposed in this research.

Additionally, online surveys conducted on different repositories would reveal user demographics, where users learned about these titles, and how they are going to interact with these titles.

Lastly, more studies need to be conducted in different countries and on different platforms. The findings should be triangulated with interviews. It is also necessary to understand why some people do not read OA monographs in digital formats or why they do not use online platforms so that these barriers can be overcome and these people can be included in scholarly communication in the digital landscape.

Appendix – Glossary

Altmetrics: Metrics that are concerned with the influence of any subject through social media using indicators of visibility and awareness such as mentions (Galligan & Dyas-Correia, 2013; Holmberg, 2014).

ACLS: The American Council of Learned Societies (http://www.acls.org/), founded in 1919, is a private, non-profit federation of 75 American scholarly organizations. It is the preeminent representative of American scholarship in the humanities and related social sciences.

Application Programming Interface (API): An API is a software intermediary that allows two applications to talk to each other. In other words, an API is the messenger that delivers the user's request to the relevant provider and delivers the provider's response back to the user. These response messages are typically expressed in the JSON or XML formats.

Bibliometrics: Bibliometrics is the statistical analysis of academic publications, such as monographs or articles.

Bielefeld Academic Search Engine (BASE): BASE is an OA academic search engine created by Bielefeld University Library in Germany.

Bounce rate: The percentage of visitors entering or landing on a website and leaving without continuing to another page on the site.

Checksum: A checksum is a value used to verify the integrity of a file or a data transfer. These can use algorithms, including md5 or sha, amongst others.

Citation vs. Reference: These are distinct terms. If a paper R contains a bibliographic note using and describing paper C, then R contains a reference to C and C has a citation from R (Solla Price, 1986). Thus, reference is a backward-looking concept, while citation is a forward-looking one (Egghe & Rousseau, 1990).

Citation index: Citation indexes track references that authors include in the reference lists of their publications. They provide a means of searching for and

analysing scholarly literature that is not possible using simple search engines. There are three main citation indexes: WoS, Scopus and Google Scholar.

Cookie: A cookie is a text file that a web browser stores on a user's machine. Websites use cookies for authentication, storing website information/preferences or other browsing information, and anything else that can help the web browser while accessing web servers.

COUNTER (or Project COUNTER): COUNTER is an international non-profit membership organization of libraries, publishers, and vendors. These members are continually developing the Code of Practice, which is a standard designed to count the usage of electronic resources.

Digital repositories: Digital repositories are used to store and disseminate scholarly information such as digital collections of books, papers, theses, media, and other works.

Directory of Open Access Books (DOAB): The DOAB is a discovery service for OA monographs maintained by the OAPEN Foundation and based at the National Library of the Netherlands.

DOI: A Digital Object Identifier or DOI is a string of numbers, letters and symbols used to permanently identify a book, scientific paper, song, image, or something else and link to it on the web.

Domain name: A domain name is an address that people use on the internet, whether for websites or for email. It is a string of characters which usually spells out a word or the name of a company, organization or person. For the URL http://ccat.curtin.edu.au/about-us.html the domain name is curtin.edu.au.

Dublin Core (DC): Dublin Core is an initiative to create a digital "library card catalogue" for the Web. Dublin Core is made up of 15 metadata (data that describe data) elements that offer expanded cataloguing information and improved document indexing for search engine programs.

Extensible Markup Language (XML): XML is a data-interchange format. Data in this format are self-describing or self-defining, meaning that the structure of the data is embedded within the data. Thus, when the data arrive, there is no need to pre-build the structure to store the data.

EZproxy: EZproxy is a web proxy server used by libraries to give access from outside the library's computer network to restricted-access websites that authenticate users by IP address.

Field normalisation: Field normalisation is the process of benchmarking monographs against other monographs within the same subject field.

Geocoding: Geocoding is the process of converting addresses (such as street addresses) into geographic coordinates (such as latitude and longitude), which can be used to place markers on a map or position the map. Reverse geocoding is the process of converting geographic coordinates into a human-readable address.

Geolocation: Geolocation is the identification or estimation of real-world geographic location.

High Integration of Research Monographs in the European Open Science (HIRMEOS): The HIRMEOS project aims to prototype innovative services for monographs in support of Open Science infrastructure by providing additional data, links and interactions to the documents. At the same time, they aim to pave the way for new tools for research assessment, which remains a major challenge in the humanities and social sciences.

Hostname: A hostname is the unique name given to a computer connected to the Internet. For example, the hostname for a website is ccat.curtin.edu.au. The first part is the local name, which in this case is ccat.

Institute for Scientific Information (ISI): The ISI is a provider of bibliographic database services. It maintains citation databases covering thousands of academic journals, including a continuation of its longtime print-based indexing service, the Science Citation Index (SCI), as well as the Social Sciences Citation Index (SSCI) and the Arts and Humanities Citation Index (AHCI). All of these are available via ISI's Web of Knowledge database service.

IP address: An Internet Protocol address (IP address) is a unique number assigned to all devices (such as computers, tablets, or phones) when they connect to the Internet.

ISBN: The International Standard Book Number (ISBN) is a unique numeric commercial book identifier.

JavaScript Object Notation (JSON): JSON is a lightweight data-interchange format. It is easy for humans to read and write and easy for machines to parse and generate. It is also faster to parse than XML.

Knowledge Unlatched (KU) – http://knowledgeunlatched.org/: Established in 2012, KU is a London-based not-for-profit company that coordinates library support and funding for OA scholarly books.

Library catalogue: A library catalogue is a register of all bibliographic items found in a library or group of libraries, such as a network of libraries at several locations.

Mertonian norms (the ethos of modern science): A set of norms and values of sciences are supposed to be built. These norms are universalism, communism, disinterestedness, and organised scepticism.

Metadata: Metadata summarize basic information about data, which can make finding and working with particular instances of data easier. Author, date created, date modified and file size are examples of basic document metadata. The ability to filter through this metadata makes it easier for users to locate specific documents.

Monograph: In academia, monographs are defined as specialist books, usually written by a single author on a single subject. In contrast to textbooks, which survey the state of knowledge in a field, monographs' main purpose is to present primary research and original scholarship. Monographs are most commonly published within the humanities and social sciences, rather than the hard sciences.

MOOC: A massive open online course (MOOC) is an online course that is freely accessible and allows unlimited participation.

OpenAIRE: OpenAIRE is a network of OA repositories, archives, and journals that support OA policies.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): The OAI-PMH specifies how metadata are structured and presented for harvesting by external services. OAI-PMH metadata are encoded in extensible markup language (XML) format.

Page tagging: Page tagging refers to the implementation of tags in the existing HTML code of a given web presence. These markings help to analyze users' behavior when they are moving between two page views.

Page views: The total number of pages that have been viewed. Repeated views of a single page are also counted. This means that if a user navigates to a different page and then returns to the same page, a second page view is recorded.

Proxy server: A proxy server is a server (a computer system or application) that acts as an intermediary for requests from clients seeking resources from other servers. In relation to this study, libraries use proxy servers to give users access from outside the library's computer network to restricted-access websites that authenticate users by IP address.

Public Library of Science (PLOS): PLOS is a nonprofit, OA science, technology and medicine publisher, innovator and advocacy organization with a library of OA journals and other scientific literature under an open content license.

Referrals: Referrals in Google Analytics are sites that "refer" visitors to another website by providing a link to the site. In most cases, this category excludes advertising visits, as well as organic searches.

Salami publishing: Salami publication or segmented publication is a distinct form of redundant publication, which is usually characterized by similarity of hypotheses, methodologies or results but not similarity of text (Šupak Smolčić, 2013).

Scientometrics: The science of measuring and analysing science.

Search engine: A website such as Google, Bing or Yahoo that assists the user in finding other web pages.

Second-level domain (SLD or 2LD): An SLD is a domain that is directly below a top-level domain (TLD). Examples of SLDs include .edu.au, .ac.uk (educational facilities); .com.au, .co.uk (commercial businesses); and .gov.au, .gov.uk (government agencies).

Semiotics: Semiotics is the systematic scholarly analysis of sign systems (Wouters, 2016).

Session: A session or visit is a unit of measurement of a user's actions performed within a particular period or in relation to the completion of a task.

Social constructivism: Social constructivism is a sociological theory of knowledge according to which human development is socially situated and knowledge is constructed through interaction with others.

Software as a service (SaaS): SaaS is a software distribution model in which a third-party provider hosts applications and makes them available to customers over the Internet. Some examples of Saas include Office365, Google Apps and Netflix.

Top-level domain (TLD): TLD refers to the last segment of a domain name, or the part that follows immediately after period. TLDs are classified into two categories: generic TLDs (gTLD) and country-code TLDs (ccTLD). Examples of some common TLDs include .com (commercial businesses), .org (organizations), .net (network organizations), .gov (U.S. government agencies), .edu (educational facilities like universities), .ca (Canada), and .au (Australia).

Unique page view: Unique page views refer to the number of sessions in which a specified page was viewed one or more times. Since the present study is concerned with each session, because it generally represents one visit, unique page views are used to count the number of separate visits to a page.

URL: "URL" stands for Uniform Resource Locator. A URL is a (mostly) human-readable string that uniquely identifies a resource (i.e., an asset, file or piece of content) on the Internet.

Web log file: A web log file is a log file automatically created and maintained by a web server. Every "hit" to a website, including each view of an HTML document, image or other object, is logged. The format of a raw web log file essentially contains one line of text for each hit on a website. This text contains information about who was visiting the site, where they came from, and exactly what they were doing on the website.

Web page: A web page is a document, commonly written in HyperText Markup Language (HTML), which is accessible through the Internet or another network using an Internet browser. A web page is accessed by entering a URL address and may contain text, graphics, and hyperlinks to other web pages and files.

Web server: A computer that hosts a website on the Internet.

Webometric Analyst: Webometric Analyst is a free software program that uses URL citations or title mentions to produce network diagrams, link impact reports, and web environment networks. It mainly uses Bing's API.

Webometrics: Webometrics aims to measure the impact of a research object across the web by examining numbers and types of hyperlinks, and employing bibliometrics approaches to examine usage patterns (Almind & Ingwersen, 1997).

Website: A collection of web pages that are grouped together and usually connected in various ways, typically identified with a common domain name and published on at least one web server.

References

Adams, J. S. (1965). Inequity in social exchange. *Advances in experimental social psychology* (Vol. 2, pp. 267–299). Elsevier.

Adema, J. (2010). Overview of open access models for ebooks in the humanities and social sciences. OAPEN, Amsterdam.

Adie, E. (2014). Attention! A study of open access vs non-open access articles. Figshare.

Adie, E., & Roe, W. (2013). Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, *26*(1), 11–17.

Alaimo, K. (2018). Twitter's Misguided Barriers for Researchers. Retrieved April 25, 2019, from https://www.bloomberg.com/opinion/articles/2018-10-16/twitter-s-barriers-for-academic-researchers-are-misguided

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLOS one*, *4*(6), e5910.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to'webometrics. *Journal of documentation*, *53*(4), 404–426.

Alperin, J. P., Gomez, C. J., & Haustein, S. (2018). Identifying diffusion patterns of research articles on Twitter: A case study of online engagement with open access articles. *Public Understanding of Science*, *28*(1), 2–18. https://doi.org/10.1177/0963662518761733

Altmetric.com. (2017). About Altmetric and the Altmetric Attention Score. Retrieved from https://help.altmetric.com/support/solutions/articles/6000059309-about-altmetric-and-the-altmetric-attention-score

Altmetric.com. (2018). How is the Altmetric Attention Score calculated?. Retrieved from https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-score-calculated-

Amichai-Hamburger, Y. (2002). Internet and personality. *Computers in Human Behavior*, *18*(1), 1–10.

Amichai–Hamburger, Y., Lamdan, N., Madiel, R., & Hayat, T. (2008). Personality characteristics of Wikipedia members. *CyberPsychology & Behavior*, *11*(6), 679–681.

Armbruster, C., & Romary, L. (2010). Comparing repository types: challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication. *International Journal of Digital Library Systems (IJDLS)*, *1*(4), 61-73.Background to OAPEN-UK. (n.d.). Retrieved from http://oapen-uk.jiscebooks.org/overview/background-to-oapen-uk/

Bar-Ilan, J. (2014). Evaluating the individual researcher–adding an altmetric perspective. *Research trends*, *37*, 31–32.Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Bhattacharya, C. B., & Sen, S. (2003). Consumer-company identification: A framework for understanding consumers' relationships with companies. *Journal of marketing*, *67*(2), 76–88.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American society for information science and technology*, *55*(14), 1216–1227.

Blanchard, A. L. (2008). Testing a model of sense of virtual community. *Computers in Human Behavior*, *24*(5), 2107–2123.

Bollen, J., Van de Sompel, H., Smith, J. A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management, 41*(6), 1419–1440. https://doi.org/10.1016/j.ipm.2005.03.024

Bonasio, A. (2013). Mendeley has 2.5 million users. *Mendeley Blog*. Retrieved from https://blog.mendeley.com/2013/09/03/mendeley-has-2-5-million-users/

Bonn, M. (2015). Maximizing the benefits of open access: Strategies for enhancing the discovery of open access content. *College & Research Libraries News*, *76*(9), 491–494.

Bornmann, L. (2014a). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of informetrics*, *8*(4), 895–903.

Bornmann, L. (2014b). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics*, *8*(4), 935–950.

Bornmann, L. (2015). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics*, *103*(3), 1123–1144.

Bornmann, L. (2016). Scientific revolution in scientometrics: the broadening of impact from citation to societal. *Theories of informetrics and scholarly communication*, 347–359.

Bornmann, L., & Haunschild, R. (2017). Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact? *Scientometrics*, *110*(2), 937–943.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

Borra, E., & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, *66*(3), 262–278.

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Information (International Social Science Council)*, *14*(6), 19–47.

Bourdieu, P. (1986). The forms of capital Handbook of theory and research for the sociology of education (pp. 241–258). New York: Greenwood.

Bruns, A., & Burgess, J. (2016). Methodological innovation in precarious spaces: The case of Twitter. *Digital Methods for Social Science* (pp. 17–33). Springer.

CAUL. (2017). Repository Statistics. Retrieved from http://archive2010.caul.edu.au/caul-programs/research/repository-services/repository-manager-tools/repository-statistics

Chen, G. M. (2011). Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. *Computers in Human Behavior*, *27*(2), 755–762.

Chen, P.-Y., Wu, S., & Yoon, J. (2004). The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings*, 58.

Cheung, C. M., & Lee, M. K. (2010). A theoretical model of intentional social action in online social networks. *Decision support systems*, *49*(1), 24–30.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, *43*(3), 345–354.

Chiu, C.-M., Hsu, M.-H., & Wang, E. T. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision support systems*, *42*(3), 1872–1888.

Chiu, C.-M., Wang, E. T., Shih, F.-J., & Fan, Y.-W. (2011). Understanding knowledge sharing in virtual communities: An integration of expectancy disconfirmation and justice theories. *Online Information Review*, *35*(1), 134–153.

Chu, H., & Krichel, T. (2007). Downloads vs. citations in economics: Relationships, contributing factors and beyond. Paper presented at the *Proceedings of ISSI 2007 - 11th International Conference of the International Society for Scientometrics and Informetrics*, pp. 207–215,

Clarivate. (2018). Web of Science platform: Web of Science: Summary of Coverage. Retrieved from https://clarivate.libguides.com/webofscienceplatform/coverage

Clarke, M. (2009). The Strength of Weak Ties: Why Twitter Matters in Scholarly Communication. *The Scholarly Kitchen*. Retrieved from

https://scholarlykitchen.sspnet.org/2009/08/12/the-strength-of-weak-ties-why-twitter-matters-in-scholarly-communication/

Collins, E., & Milloy, C. (2016). OAPEN-UK final report: A five-year study into open access monograph publishing in the humanities and social sciences.

Cooper, A., & Postel, J. (1993). The US domain; Request for comments: 1480. Marina del Rey, CA: University of Southern California's Information Sciences Institute. Retrieved May, 28, 2007.

Costas, R., Zahedi, Z., & Wouters, P. (2015a). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019.

Costas, R., Zahedi, Z., & Wouters, P. (2015b). The thematic orientation of publications mentioned on social media: Large-scale disciplinary comparison of social media metrics with citations. *Aslib Journal of Information Management*, *67*(3), 260–288.

Cronin, B. (1981). The need for a theory of citing. *Journal of documentation*, *37*(1), 16–24.

Cronin, B. (2000). Semiotics and evaluative bibliometrics. *Journal of Documentation*, *56*(4), 440–453.

Cronin, B. (2016). Theories of Informetrics and Scholarly Communication. In C. Sugimoto (Ed.), (pp. 13–19). De Gruyter.

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the Web. *Journal of the American Society for Information Science*, *49*(14), 1319–1328.

Cropanzano, R., & Mitchell, M. S. (2005). Social exchange theory: An interdisciplinary review. *Journal of management*, *31*(6), 874–900.

Crossick, G. (2015). Monographs and Open Access A report to HEFCE.

Cutler, N. E., & Danowski, J. A. (1980). Process gratification in aging cohorts. *Journalism Quarterly*, *57*(2), 269–276.

Czarniawska, B. (1997). *A narrative approach to organization studies* (Vol. 43). Sage Publications.

Davenport, T. H., & Beck, J. C. (2001). *The attention economy: Understanding the new currency of business*. Harvard Business Press.

Delgado López-Cózar, E., Orduña-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a Data Source for Research Assessment. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 95–127). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_4

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, *41*(1), 417–440.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier Science Publishers.

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of computer-mediated communication*, *11*(2), 415–441.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. Journal of computer-mediated communication, 12(4), 1143-1168.

Elsevier. (2016). Scopus: content coverage guide. Amsterdam: Elsevier.

Ely, R. J. (1994). The effects of organizational demographics and social identity on relationships among professional women. *Administrative science quarterly*, 203-238.

Emery, C. (2015). Over 35,000 downloads of 28 HSS monographs in 167 countries. Retrieved from http://www.knowledgeunlatched.org/2015/09/usage-stats-5/

Eve, M. P. (2014). *Open access and the humanities*. Cambridge University Press.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research*, *13*(4), e123.

Fenner, M. (2014). Altmetrics and other novel measures for scientific impact. *Opening science* (pp. 179–189). Springer.

Ferwerda, E., Snijder, R., & Adema, J. (2013). *OAPEN-NL: a project exploring open Access monograph publishing in the Netherlands; final report*. Oapen Foundation.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291–313.

Forte, A., & Bruckman, A. (2005). Why do people write for Wikipedia? Incentives to contribute to open–content publishing. *Proc. of GROUP*, *5*, 6–9.

Fox, S., & Lenhart, A. (2006). Bloggers: A portrait of the internet's new storytellers. *Pew Internet & American Life Project*, *19*.

Fredette, J. (2015). The Islamic State's attacks on Paris were attacks on Muslims, too. Retrieved January 10, 2018, from https://www.washingtonpost.com/news/monkey-cage/wp/2015/11/16/the-islamic-states-attacks-on-paris-were-attacks-on-muslims-too/

Fung, H. P. (2015). How can we distinguish between mediator and moderator variable, theoretically? *Researchgate discussion page*. Retrieved from https://www.researchgate.net/post/How_can_we_distinguish_between_mediator_and_moderator_variable_theoretically

Galligan, F., & Dyas-Correia, S. (2013). Altmetrics: Rethinking the way we measure. *Serials review*, *39*(1), 56–61.

Gardiner, E., & Musto, R. G. (2005). Proceedings from the 2004 APA Panel: Electronic Publication and the Classics Profession. *Syllecta Classica*, *16*, 221–229.

Garfield, E. (1964). Science Citation Index"-A New Dimension in Indexing. *Science*, *144*(3619), 649–654.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, *1*(4), 359–375.

Garfield, E., & others. (1965). Can citation indexing be automated. *Statistical association methods for mechanized documentation, symposium proceedings* (Vol. 269, pp. 189–192). National Bureau of Standards, Miscellaneous Publication 269, Washington, DC.

Gilbert, G. N. (1977). Referencing as persuasion. *Social studies of science*, *7*(1), 113–122.

Giles, J. (2005). Internet encyclopaedias go head to head. Nature Publishing Group.

Gilpin, D. (2011). Working the Twittersphere: Microblogging as professional identity construction. *A networked self: Identity, community and culture on social network sites. New York: Routledge*, 232–250.

Gingras, Y. (2014). Criteria for evaluating indicators. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact, 110-125.

Goffman, E., & others. (1978). *The presentation of self in everyday life*. Harmondsworth.

Goodreads. (n.d.). About Goodreads. Retrieved from https://www.goodreads.com/about/us

Google. (2009). Back to Basics: "not set" Entries. *Google Analytics Solutions*. Retrieved from https://analytics.googleblog.com/2009/08/back-to-basics-not-set-entries.html

Google. (2011). Facts about Google and Competition. Archived from the original on 4 November 2011. Retrieved from https://web.archive.org/web/20111104131332/https://www.google.com/competition/howgooglesearchworks.html

Gosling, S. D., Gaddis, S., Vazire, S., & others. (2007). Personality impressions based on facebook profiles. *ICWSM*, *7*, 1–4.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 1360–1380.

Groth, P., & Gurney, T. (2010). Studying scientific discourse on the Web using bibliometrics: A chemistry blogging case study.

Hagstrom, W. (1982). Gift Giving as an Organizing Principle in Science. In B. Barnes & D. O. Edge (Eds.), *Science in Context: Readings in the Sociology of Science* (pp. 21–34). MIT Press.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. *Scientometrics*, *101*(2), 1419–1430.

Harzing, A.-W. (2017). Google Scholar is a serious alternative to Web of Science. *LSE Impact Blog*.

Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787–804.

Haustein, S. (2015). Scientific Interactions and Research Evaluation: From Bibliometrics to Altmetrics. *ISI* (pp. 36–42).

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PloS one, 10*(3). e0127830. https://doi.org/10.1371/journal.pone.0127830

Haustein, S., Bowman, T. D., & Costas, R. (2015). Interpreting“ altmetrics”: Viewing acts on social media through the lens of citation and social theories.In C. R. Sugimoto (Ed.), Theories of Informetrics and Scholarly Communication (pp. 372-405). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110308464-022

Haustein, S., & Larivière, V. (2014). Mendeley as a source of readership by students and postdocs? Evaluating article usage by academic status. In *IATUL Conference, Espoo, Finland, June 2-5 2014*. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2033&context=iatul

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014a). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics, 101*(2), 1145–1163. https://doi.org/10.1007/s11192-013-1221-3

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014b). Tweeting biomedicine: An analysis of tweets and citations in the biomedical

literature. *Journal of the Association for Information Science and Technology*, *65*(4), 656–669.

Haustein, S., Sugimoto, C. R., & Larivière, V. (2015). Social media in scholarly communication. *Aslib Journal of Information Management*, *67*(3), 1–14.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, *520*, 429–431.

HIRMEOS. (2017). Metrics Services Specification. Retrieved from http://www.hirmeos.eu/wp-content/uploads/2017/11/HI61-Metrics_Service_technical_specification-final.pdf

Holmberg, K. (2014). The meaning of altmetrics. *IATUL annual conference proceedings*, (Vol. 35, pp. 1–11). Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=llf&AN=97787073&site=ehost-live

Hossain, L., & Silva, A. de. (2009). Exploring user acceptance of technology using social networks. *The Journal of High Technology Management Research*, *20*(1), 1–18.

Hsu, C.-L., & Lin, J. C.-C. (2008). Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. *Information & management*, *45*(1), 65–74.

Huysman, M., Wulf, V., & others. (2004). *Social capital and information technology*. Cambridge, MA: MIT Press.

ImpactStory. (2012). A new framework for altmetrics. *ImpactStory blog*. Available at: http://blog.impactstory.org/2012/09/14/31524247207/.

Introduction: Defining Research Impacts. (n.d.).*The Impact Blog*. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/introduction/

Ip, R. K. F., & Wagner, C. (2008). Weblogging: A study of social computing and its impact on organizations. *Decision Support Systems*, *45*(2), 242–250.

IRUS-UK. (n.d.). IRUS-UK. Retrieved from http://irus.mimas.ac.uk/about/

ISO, I. (2008). 32000-1: 2008, Document Management–Portable Document Format–Part 1: PDF 1.7. International Organization for Standardization, Geneva, Switzerland.

Jansen, J. (2010). Online Product Research. *Pew Research Center Internet & Technology Report*. Retrieved from http://www.pewinternet.org/2010/09/29/online-product-research/

JISC. (n.d.). Who we are and what we do. Retrieved from https://www.jisc.ac.uk/about/who-we-are-and-what-we-do

JSTOR. (2016). Topic Cards and You, Retrieved from http://support.jstor.org/additional-resources-student-and-faculty/2017/4/7/topic-cards-and-you

Kahle, B. (2018). Digital Books on archive.org. *Internet Archive Blogs*. Retrieved from https://blog.archive.org/2018/01/24/digital-books-on-archive-org/

Kimmons, R. M. (2011). Understanding collaboration in Wikipedia. *First Monday*, *16*(12).

Kinsley, S. (2015). A political economy of Twitter data? Conducting research with proprietary data is neither easy nor free. *Impact of Social Sciences Blog*. Retrieved from https://blogs.lse.ac.uk/usappblog/2015/01/03/a-political-economy-of-twitter-data-conducting-research-with-proprietary-data-is-neither-easy-nor-free/

Kjellberg, S. (2010). I am a blogging researcher: Motivations for blogging in a scholarly context. *First Monday*, *15*(8).

Kling, R. and Callahan, E. (2003), Electronic journals, the Internet, and scholarly communication. *Ann. Rev. Info. Sci. Tech., 37*: 127-177. doi:10.1002/aris.1440370105

Knowledge Unlatched. (2017). Knowledge Unlatched Releases Geolocational Data for Pilot and Round 2 Collections. Retrieved from http://www.knowledgeunlatched.org/2017/05/knowledge-unlatched-releases-geolocation-data-for-pilot-and-round-2-collections/

Konkiel, S., Sugimoto, C. R., & Williams, S. (2016). What constitutes valuable scholarship? The use of altmetrics in promotion and tenure. Impact of Social Sciences Blog.

Kortelainen, T., & Katvala, M. (2012). Everything is plentiful—Except attention". Attention data of scientific journals on social web tools. *Journal of Informetrics*, *6*(4), 661–668.

Kousha, K., & Thelwall, M. (2014). Web Impact Metrics for Research Assessment. In B. Cronin & C. Sugimoto (Eds.), *Beyond bibliometrics: Harnessing multi-dimensional indicators of performance*. Cambridge, MA: MIT Press.

Kousha, K., & Thelwall, M. (2016). Can Amazon. com reviews help to assess the wider impacts of books? *Journal of the Association for Information Science and Technology*, *67*(3), 566–581.

Kousha, K., & Thelwall, M. (2017). Are wikipedia citations important evidence of the impact of scholarly articles and books? Journal of the Association for Information Science and Technology, 68(3), 762–779. https://doi.org/10.1002/asi.23694

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the Association for Information Science and Technology*, *62*(11), 2147–2164.

Kwon, O., & Wen, Y. (2010). An empirical study of the factors affecting social network service use. *Computers in human behavior*, *26*(2), 254–263.

Lazarsfeld, P. F. (1958). Evidence and inference in social research. *Daedalus, 87*(4), 99-130.

Leydesdorff, L., Bornmann, L., Comins, J. A., & Milojevi, S. (2016). Citations: Indicators of quality? The impact fallacy. *Frontiers in Research Metrics and Analytics*, *1*, 1.

Li, X., Thelwall, M., & Giustini, D. (2011). Validating online reference managers for scholarly impact measurement. *Scientometrics*, *91*(2), 461–471.

Lin, J. (2012). A case study in anti-gaming mechanisms for altmetrics: PLOS ALMs and DataTrust. *paper, altmetrics12 ACM Web Science Conference, Evanston, IL*.

Lin, J., & Fenner, M. (2013). Altmetrics in evolution: defining and redefining the ontology of article-level metrics. *Information Standards Quarterly*, *25*(2), 20.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435–444.

Mahrt, M., Weller, K., & Peters, I. (2014). Twitter in scholarly communication. *Twitter and society*, 399–410.

Marashi, S.-A., Hosseini-Nami, S. M. A., Alishah, K., Hadi, M., Karimi, A., Hosseinian, S., Fard, R. R., et al. (2013). Impact of Wikipedia on citation trends. *EXCLI journal*, *12*, 15.

Marshall, T. C., Ferenczi, N., Lefringhausen, K., Hill, S., & Deng, J. (2020). Intellectual, narcissistic, or Machiavellian? How Twitter users differ from Facebook-only users, why they use Twitter, and what they tweet about. *Psychology of Popular Media, 9*(1), 14–30. https://doi.org/10.1037/ppm0000209

Mas-Bleda, A., Thelwall, M., Kousha, K., & Aguillo, I. F. (2014). Do highly cited researchers successfully use the social web? Scientometrics, 101(1), 337–356. https://doi.org/10.1007/s11192-014-1345-0

McCollough, A. (2017). Does It Make a Sound: Are Open Access Monographs Discoverable in Library Catalogs? *portal: Libraries and the Academy*, *17*(1), 179–194.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Merton, R. K. (2000). On the Garfield input to the sociology of science: A retrospective collage. In B. C. Helen Barsky Atkins (Ed.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 435–448). ASIS.

Meyers, J. (2014). How to Search Public Posts on Facebook. Retrieved from https://digiwonk.gadgethacks.com/how-to/search-public-posts-facebook-0155649/

Millington, P. (2006). OpenDOAR - Home Page - Directory of Open Access Repositories. Retrieved from http://www.opendoar.org/

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology, 56*(10), 1088–1097. https://doi.org/10.1002/asi.20200

Moed, H. F. (2015). Altmetrics as traces of the computerization of the research process. *arXiv preprint arXiv:1510.05131*.

Moed, H. F., & Halevi, G. (2015). Multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology*, *66*(10), 1988–2002.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, *65*(8), 1627–1638.

Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015). Who reads research articles? An altmetrics analysis of Mendeley user categories. *Journal of the Association for Information Science and Technology*, *66*(9), 1832–1846.

Mohammadi, E., Thelwall, M., & Kousha, K. (2015). Can Mendeley bookmarks reflect readership? A survey of user motivations. *Journal of the Association for Information Science and Technology, 67*(5), pp.1198-1209.

Mohammed, S., Ferzandi, L., & Hamilton, K. (2010). Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management*, *36*(4), 876–910.

Montgomery, L. (2013). Metrics Challenges for Monographs. Retrieved from http://www.knowledgeunlatched.org/2013/04/metrics-challenges-for-monographs/

Montgomery, L., Ozaygen, A., Pinter, F., & Saunders, N. (2017). Exploring the Uses of Open Access Books via the JSTOR Platform.

Montgomery, L., Neylon, C., Ozaygen, A., & Leaver, T. (2018). Getting the best out of data for open access monograph presses: A case study of UCL Press. *Learned Publishing*.

Moser, A., & Korstjens, I. (2018). Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. *European Journal of General Practice, 24*(1), 9–18. https://doi.org/10.1080/13814788.2017.1375091

Nahapiet, J., & Ghoshal, S. (2000). Social capital, intellectual capital, and the organizational advantage. *Knowledge and social capital* (pp. 119–157). Elsevier.

Nederhof, A. J. (2011). A bibliometric study of productivity and impact of modern language and literature research. *Research Evaluation*, *20*(2), 117–129.

Nelson, C. (1997). Superstars. *Academe*, *83*(1), 38–54.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Sage Publications Inc., Thousand Oaks, CA.

Neylon, C. (2014). Altmetrics: What are they good for. *PLOS Opens*. Retrieved from http://blogs.plos.org/opens/2014/10/03/altmetrics-what-are-they-good-for/

Neylon, C., Montgomery, L., Ozaygen, A., Pinter, F., & Saunders, N. (2018). The Visibility of Open Access Monographs in a European Context: A Report Prepared by Knowledge Unlatched Research.

Ngai, E. W., Moon, K. K., Lam, S. S., Chin, E. S., & Tao, S. S. (2015). Social media models, technologies, and applications: an academic review and case study. *Industrial Management & Data Systems*, *115*(5), 769–802.

Ngai, E. W., Tao, S. S., & Moon, K. K. (2015). Social media research: Theories, constructs, and conceptual frameworks. *International Journal of Information Management*, *35*(1), 33–44.

Nicolaisen, J. (2007). Citation analysis. *Annual review of information science and technology*, *41*(1), 609–641.

Nielsen, F. Å. (2007). Scientific citations in Wikipedia. *arXiv preprint arXiv:0705.2106*.

NISO. (2014). NISO Altmetrics Standards Project White Paper. Retrieved from https://groups.niso.org/apps/group_public/download.php/13295/niso_altmetrics_white_paper_draft_v4.pdf

Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. Libri, 55(4), 170-180.

Nov, O. (2007). What motivates wikipedians? *Communications of the ACM*, *50*(11), 60–64.

OAPEN. (n.d). Annual Report. Retrieved from https://oapen.org/content/about-annual-report-2015

OAPEN-UK. (2014). OAPEN-UK Librarian Survey. Retrieved from http://oapen-uk.jiscebooks.org/research-findings/librarian-survey/

Obrien, P., Arlitsch, K., Sterman, L., Mixter, J., Wheeler, J., & Borda, S. (2016). Undercounting file downloads from institutional repositories. *Journal of Library Administration*, *56*(7), 854-874.

OpenAIRE. (n.d.). Project factsheets. Retrieved from https://www.openaire.eu/project-factsheets

OpenEdition (n.d). The OpenEdition Freemium programme. Retrieved from http://www.openedition.org/14043?lang=en

OPERAS. (2018). About. Retrieved from https://operas.hypotheses.org/aboutoperas

Orduña-Malea, E., Ayllón, J. M., Martɯn-Martɯn, A., & López-Cózar, E. D. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931–949.

Padula, D., & Williams, C. (2015). Applied Altmetrics: How university presses, academic publishing services and institutional repositories benefit. *Impact of Social Sciences Blog*.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (*Vol. 10*, No. 2010, pp. 1320-1326).

Piwowar, H. (2012). A new framework for altmetrics. *ImpactStory blog*. Retrieved from http://blog.impactstory.org/31524247207/

Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, *493*(7431), 159.

Price, G. (2015). Content from Bielefeld University's BASE Database Now Searchable in EBSCO Discovery Service. Retrieved from

https://www.infodocket.com/2015/12/07/content-from-bielefeld-universitys-base-database-now-searchable-in-ebsco-discovery-service/

Priem, J. (2010). I like the term #articlelevelmetrics, but it fails to imply *diversity* of measures. Lately, I'm liking #altmetrics. 21 September 2010, 3:28 a.m. Tweet.

Priem, J. (2014). Altmetrics. *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, 263–88.

Priem, J., Groth, P., & Taraborelli, D. (2012). The altmetrics collection. *PLOS one*, *7*(11), e48753.

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv preprint arXiv:1203.4745*.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto.

Project COUNTER (2012). The COUNTER Code of Practice for e-Resources: Release 4.

Gorraiz, J., Purnell, P. J., & Glänzel, W. (2013). Opportunities for and limitations of the B ook C itation I ndex. *Journal of the American Society for Information Science and Technology*, *64*(7), 1388-1398.

Puschmann, C., & Mahrt, M. (2012). Scholarly blogging: A new form of publishing or science journalism 2.0. *Science and the Internet*, 171–182.

Qantcast. (2018). Goodreads.com. In Quantcast [Audience Insight]. Retrieved from https://www.quantcast.com/goodreads.com?country=US

Ravetz, J. R. (1973). *Scientific knowledge and its social problems*. Transaction publishers.

ROAR. (n.d.). Welcome to the Registry of Open Access Repositories. Retrieved from http://roar.eprints.org/

Rui, H., & Whinston, A. (2012). Information or attention? An empirical study of user contribution on Twitter. *Information Systems and e-Business Management*, *10*(3), 309–324.

Ryan, J., Avelar, I., Fleissner, J., Lashmet, D. E., Miller, J. H., Pike, K. H., Sitter, J., et al. (2002). The Future of Scholarly Publishing: MLA Ad Hoc Committee on the Future of Scholarly Publishing. *Profession*, 172–186.

Sammer, T., & Back, A. (2011). Towards Microblogging Success Factors: An Empirical Survey on Twitter Usage of Austrian Universities. *MCIS 2011 Proceedings*, 40.

Schubert, A., & Braun, T. (1996). Cross-field normalisation of scientometric indicators. *Scientometrics*, *36*(3), 311–324.

Sharma, H. (2018). PaperHive: Tool review and tutorial for researchers. *Editage Insights*. Retrieved from https://www.editage.com/insights/paperhive-tool-review-and-tutorial-for-researchers

Shema, H., Bar-Ilan, J., & Thelwall, M. (2012). Research blogs and the discussion of scholarly information. *PLOS one*, *7*(5), e35869.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2014). Scholarly blogs are a promising altmetric source. *Research Trends*, *37*, 11–13.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2015). How is research blogged? A content analysis approach. *Journal of the Association for Information Science and Technology*, *66*(6), 1136–1149.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PLOS one*, *7*(11), e47523.

Simon, H. A. (1971). Designing organizations for an information-rich world.

Small, H. G. (1978). Cited documents as concept symbols. *Social studies of science*, *8*(3), 327–340.

Small, H. (2004). On the shoulders of Robert Merton: Towards a normative theory of citation. Scientometrics, 60(1), 71-79.

Small, H. G. (2016). Referencing as Cooperation or Competition. In C. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication* (pp. 49–71). Walter de Gruyter GmbH & Co KG.

Smith, A., & Anderson, M. (2016). Online Shopping and E-Commerce. *Pew Research Center Internet & Technology Report*. Retrieved from http://www.pewinternet.org/2016/12/19/online-shopping-and-e-commerce/

Smith, L. C. (1981). Citation analysis.

Snijder, R. (2010). The profits of free books: an experiment to measure the impact of open access publishing. *Learned Publishing*, *23*(4), 293–301.

Snijder, R. (2016). Revisiting an open access monograph experiment: measuring citations and tweets 5 years later. *Scientometrics*, *109*(3), 1855–1875.

Solla Price, D. J. de. (1986). *Little science, big science... and beyond*. Columbia University Press New York.

Solla Price, D. J. de (1970). Citation measures of hard science, soft science, technology, and nonscience. *Communication among scientists and engineers*, 3-22.

Springer. (2010). A Survey of E-book Usage and Perceptions at the University of Liverpool: University of Liverpool E-book Study: Part 2. Retrieved from http://static.springer.com/sgw/documents/1343310/application/pdf/V7671_Liverpool_White_Paper_Part2%5B1%5D.pdf

Steele, C. (2008). Scholarly monograph publishing in the 21st century: The future more than ever should be an open book. *Journal of Electronic Publishing*, *11*(2).

Strathern, M. (1997). 'Improving ratings': audit in the British University system. *European review*, *5*(3), 305-321.

Sugimoto, C. R. (2015). Attention is not impact" and other challenges for altmetrics. *Discover the future of research: Wiley Exchanges*. Retrieved from https://hub.wiley.com/community/exchanges/discover/blog/2015/06/23/attention-is-not-impact-and-other-challenges-for-altmetrics

Sugimoto, C. R., & Larivière, V. (2018). Measuring Research: What Everyone Needs to Know®. Oxford University Press, Oxford.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2016). Scholarly use of social media and altmetrics: a review of the literature. *Journal of the*

Association for Information Science and Technology, 68(9), 2037–2062. https://doi.org/10.1002/asi.23833

Tan, A. S. (1985). *Mass communication theories and research*. New York: Wiley.

Taylor, M. (2013). Towards a common model of citation: Some thoughts on merging altmetrics and bibliometrics. *Research Trends*, *35*, 19–22.

Taylor, M. (2013). Exploring the boundaries: How altmetrics can expand our vision of scholarly communication and social impact. *Information Standards Quarterly*, *25*(2), 27–32.

Teplitskiy, M., Lu, G., & Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, *68*(9), 2116–2127.

Terras, M. (2012). The impact of social media on the dissemination of research: Results of an experiment. *Journal of Digital Humanities*, *1*(3). Retrieved from http://journalofdigitalhumanities.org/1-3/the-impact-of-social-media-on-the-dissemination-of-research-by-melissa-terras/

The Digital Methods Initiative. (2015). TCAT: The New Twitter Modeling Tool for Visualizing Social Media Data. Retrieved from http://painepublishing.com/measurementadvisor/pretty-pictures-new-twitter-modeling-tool-can-make-social-media-data-tangible-actionable/

The Digital Methods Initiative. (n.d.). The Digital Methods Initiative - About Us. Retrieved from https://wiki.digitalmethods.net/Dmi/DmiAbout

Thelwall, M. (2001). Extracting macroscopic information from web links. *Journal of the American Society for Information Science and Technology*, *52*(13), 1157–1168.

Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. *Synthesis lectures on information concepts, retrieval, and services*, *1*(1), 1–116.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLOS one*, *8*(5), e64841.

Thelwall, M., & Sud, P. (2012). Webometric research with the Bing Search API 2.0. *Journal of Informetrics*, *6*(1), 44–52.

Thelwall, M., & Kousha, K. (2017). Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, *68*(4), 972-983.

Tinkler, J. (2015). Rather than narrow our definition of impact, we should use metrics to explore richness and diversity of outcomes. *Impact of Social Sciences Blog*. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/2015/07/28/impact-metrics-and-the-definition-of-impact-tinkler/

Tonia, T., Van Oyen, H., Berger, A., Schindler, C., & Künzli, N. (2016). If I tweet will you cite? The effect of social media exposure of articles on downloads and citations. *International journal of public health*, *61*(4), 513–520.

Torres-Salinas, D., Cabezas-Clavijo, Á., & Jiménez-Contreras, E. (2013). Altmetrics: New indicators for scientific communication in web 2.0. *arXiv*

Van Noorden, R. (2013). Brazilian citation scheme outed. *Nature*, *500*(7464), 510–511.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: What is the difference? *Journal of the American Society for Information Science and Technology*, *54*(14), 1313–1322.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391.

Waltman, L., & Costas, R. (2013). F1000 recommendations as a new data source for research evaluation: A comparison with citations. arXiv preprint. *arXiv preprint arXiv:1303.3875*.

Wang, X., Liu, C., Mao, W., & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. Scientometrics, 103(2), 555-564.

Wang, X., Wang, Z., & Xu, S. (2013). Tracing scientist's research trends realtimely. *Scientometrics*, *95*(2), 717–729.

Weissmann, J. (2013). The Simple Reason Why Goodreads Is So Valuable to Amazon. *The Atlantic*. Retrieved from https://www.theatlantic.com/business/archive/2013/04/the-simple-reason-why-goodreads-is-so-valuable-to-amazon/274548/

What outputs and sources does Altmetric track? (2018).*Altmetric Support*. Retrieved from https://help.altmetric.com/support/solutions/articles/6000060968-what-outputs-and-sources-does-altmetric-track-

Wikipedia:Service awards. (n.d.). Wikipedia:Service awards — Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Service_awards

Williams, C. (2015). Publishers: making altmetrics work for you. *Altmetric Blog*. Retrieved from http://www.altmetric.com/blog/publishers-making-altmetrics-work-for-you/

Willinsky, J. (2009). Toward the design of an open monograph press. *Journal of Electronic Publishing*, *12*(1).

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., et al. (2015). The metric tide: Report of the independent review of the role of metrics in research assessment and management.

Wouters, P. (1999). *The citation culture* (Doctoral dissertation, Universiteit van Amsterdam).

Wouters, P. (2016). Semiotics and citations. In C. Sugimoto (Ed.), *Theories of informetrics and scholarly communication* (pp. 72–92). Walter de Gruyter GmbH & Co KG.

Wouters, P., & Costas, R. (2012). *Users, narcissism and control: tracking the impact of scholarly publications in the 21st century*. SURFfoundation Utrecht.

Wu, T. (2015). What Ever Happened to Google Books. *The New Yorker*. Retrieved from https://www.newyorker.com/business/currency/what-ever-happened-to-google-books

Wu, W., & Zheng, R. (2012). The impact of word-of-mouth on book sales: review, blog or tweet? *Proceedings of the 14th Annual International Conference on Electronic Commerce* (pp. 74–75). ACM.

Xia, F., Su, X., Wang, W., Zhang, C., Ning, Z., & Lee, I. (2016). Bibliographic analysis of nature based on twitter and facebook altmetrics data. *PLOS one*, *11*(12), e0165997.

Yan, K.-K., & Gerstein, M. (2011). The spread of scientific information: insights from the web usage statistics in PLOS article-level metrics. *PLOS One*, *6*(5), e19917.

Yardi, S., Golder, S. A., & Brzozowski, M. J. (2009). Blogging at work and the corporate attention economy. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2071–2080). ACM.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of "alternative metrics" in scientific publications. *Scientometrics*, *101*(2), 1491–1513.

Zaytsev, A. (2017). 14 Million Books & 6 Million Visitors: HathiTrust Growth and Usage in 2016. *Perspectives from Hathitrust*. Retrieved from https://www.hathitrust.org/blogs/perspectives-from-hathitrust/14-million-books-6-million-visitors

Zeifman, I. (2015). 2015 Bot Traffic Report: Humans Take Back the Web, Bad Bots Not Giving Any Ground. Retrieved from https://www.incapsula.com/blog/bot-traffic-report-2015.html

Zhang, Y., & Hiltz, S. R. (2003). Factors that influence online relationship development in a knowledge sharing community. *AMCIS 2003 proceedings*, 53.

Zhong, B., Hardin, M., & Sun, T. (2011). Less effortful thinking leads to more social networking? The associations between the use of social network sites and personality traits. *Computers in Human Behavior*, *27*(3), 1265–1271.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, *74*(2), 133–148.

Zivkovic, B. (2011). What is: ResearchBlogging. org| The Network Central. *Scientific American Blog Network. Scientific American blogs. Retrieved April*, *16*, 2014.

Ziman, J. (2002). Real science: What it is and what it means. Cambridge University Press.

Zuccala, A. A., Verleysen, F. T., Cornacchia, R., & Engels, T. C. (2015). Altmetrics for the humanities: Comparing Goodreads reader ratings with citations to history books. *Aslib Journal of Information Management*, *67*(3), 320–336.

Zunde, P. (1971). Structural models of complex information sources. *Information storage and retrieval, 7*(1), 1-18.