View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Lund University Publications - Student Papers

THE OPEN ACCESS LANDSCAPE OF INSTITUTIONAL REPOSITORIES

A bibliometric analysis of Lund University Publications

Amanda Morrill

Thesis (30 higher education credits) for a master's degree (two-year) in library and information science within the ABM master's program at Lund University
Supervisor: Fredrik Åström
Year: 2015

© Amanda Morrill

Title

The Open Access Landscape of Institutional Repositories: A bibliometric analysis of Lund University Publications

Abstract

This master's thesis is a bibliometric study of the institutional repository Lund University Publishing. The study aims to investigate the role of institutional repositories in open access publishing and to examine institutional repositories' potential in describing research impact through the download statistics they provide for their open access material. The Swedish Research Council's recent report on research evaluation in Sweden and the planned development of the SwePub national research database are impetus for a closer evaluation of institutional repositories in Sweden. Although related research has approached these topics before, the current thesis will do so uniquely by combining metadata and usage data from an institutional repository with citation analysis from Web of Science to better understand the role of institutional repositories in making research output available open access. The research output data of three faculties at the large, multidisciplinary higher education institution Lund University were examined for the study. The results of the bibliometric analysis showed a trend of increased proportions of publications available open access in the institutional repository from 2008 to 2012. There were however differences between the faculties, with the Faculty of Medicine and the Faculty of Engineering showing a marked increase in open access content while the Faculties of the Humanities and Theology showed no increase. No substantial correlation was found between downloads and citations, suggesting that downloads represent a unique indicator of open access research impact, potentially reflecting use of scholarly material by a user group other than researcher colleagues. Further studies will however be needed to investigate the meaning of usage data in the form of download statistics.

Keywords: institutional repositories, open access, bibliometrics, scientometrics, download statistics

Acknowledgements

I would like to thank my supervisor Fredrik Åström, specialist in bibliometrics at Lund University Libraries, for his guidance during the writing process.

I would also like to thank the student union at the Faculties of the Humanities and Theology, whose members work hard to ensure the continued excellence of education at our faculty, with often very little thanks.

All faults in the thesis are mine and mine alone.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

APC……. Article processing charge
FE…… Faculty of Engineering
FHT……. Faculties of the Humanities and Theology
FM……... Faculty of Medicine
IR……. Institutional repository
LUP……...Lund University Publications
OA……... Open access
PPMCC……Pearson product-moment correlation coefficient
WoS…….Web of Science

LIST OF FIGURES

1. Introduction

The institutional repository (IR) has become a staple of the academic library in their mission to spread the research output of their establishments, and these repositories now play a key role in making scholarly output freely accessible to all on the Internet. This free accessibility is known as open access (OA), and IRs' role in OA can take various forms, from the IR being the main platform for digital publishing to diverse forms of parallel publishing. According to the IR Lund University Publications' (LUP) website, "The main aim of this service is to make your publication as visible and accessible on the Internet as possible" (Graffner 2014). Both LUP[1] and Sweden's national research publication database, SwePub, are scheduled for major upgrades (Swedish Research Council 2014, p. 59). SwePub, a national research database that collects data from IRs at Swedish universities, has in a newly released report been suggested to play an important role in the performance-based research allocation process in Sweden (ibid). Furthermore, the Swedish Research Council has mandated that research receiving funding from them must publish their results OA, increasing the importance of well-working IRs.

Lund University Libraries and academic libraries in general invest considerable time and effort in the maintenance of IRs, but it is not always clear to what effect. Understanding the role of IRs in OA is complicated by the fact that LUP and similar IRs have the potential to produce usage data for their OA content in the form of download statistics, but it is as yet unclear how these measurements can affect the evaluation of research impact in relation to traditional bibliometric citation analysis. Furthermore, as discussed below, the need for increasing OA scholarly publishing has been partially motivated by giving the public access to publicly funded research output, making IRs' fulfillment of their OA ambitions relevant from a democratic perspective. The aim of this thesis is thus to investigate the role of IRs in OA publishing and to examine the potential of IRs in describing research impact through the download statistics they provide for their OA material.

In the context of the burgeoning fields of IR research and research investigating alternatives to traditional citation analysis as a quantitative measurement for assessing research impact, there is a need for additional or complementary critical investigation into the role of IRs. The types of texts that IRs are comprised of need to be studied in relation to the work organization of science and how IR download statistics relate to traditional bibliometric measurements

[1] Fredrik Åström, librarian, supervisor meeting, 10th of February 2015

of research output. Given differences in communication behavior between academic subject fields and that the OA movement is a relatively new phenomenon in the history of scientific communication, it is necessary to analyze IRs and their OA role both over time and from different academic fields in order to build up a more complete picture of how the IR OA landscape can be visualized. This investigation is of particular concern given that IR metadata plays an increasingly important role in research evaluation procedures at Swedish universities, and these procedures need to be developed and refined to best evaluate research quality and impact in a way that treats all academic subject areas fairly. A review of IRs, the instruments used in these evaluation procedures, is a prerequisite for fair treatment.

The study presented in this thesis will examine the OA landscape from two perspectives, a descriptive analysis and a comparative analysis, both of which are grounded in the need to explore how the OA mission manifests itself and is fulfilled in IRs. The study will employ a quantitative bibliometric methodology, which will allow for the analysis of a large dataset, enabling the uncovering of patterns within large amounts of information. Due to limitations in the size of the thesis, one representative IR was chosen for investigation, Lund University's IR LUP. Lund University's IR is a good candidate for the purposes of the study because Lund University is a large, multidisciplinary and well-established university. Founded in 1666, the university encompasses eight faculties with a strong focus on research. In addition, the university's IR contains metadata on all research output as well as for the full-text publications it stores, which will provide a good basis for examination of open access publications in relation to the totality of research output at Lund University.

The first, descriptive part of the analysis will examine trends in the availability of publications OA by document type. In the second part of the study, the download statistics that the IR provides will be analyzed and correlated with traditional bibliometric indicators derived from citation analysis in order to understand how downloads relate to citation analysis as a research evaluation tool. The first part of the study will provide necessary information and context for the interpretation of the second part of the study. Furthermore, the publication data will be examined by faculty association to give a more complete picture of the OA scientific communication behavior. The results of both parts of the study will be used to answer the following research questions:

- What does the OA landscape look like in an IR and what can IR metadata, download statistics and citation analysis together reveal about the organization of work in scholarly research output in relation to OA publishing?

- How do the download statistics of OA publications in an IR compare with Web of Science citation analysis and how can download statistics from an IR inform our understanding of research impact?

The study will examine three of Lund University's faculties – the Faculty of Medicine, the Faculty of Engineering, and the Faculties of the Humanities and Theology – over a span of five years from 2008 to 2012.

The thesis will contribute to both the intellectual base of library and information science and its practical application. On the theoretical side, the thesis will evaluate the role of IRs in making research output available OA and expand the understanding of download statistics in IRs as a measure of research impact. On the practical side, the thesis will help academic librarians improve their work with IRs and make more well-informed policy decisions on the role of IRs in widening research impact.

1.1 Disposition

In the introduction to this thesis, the reader has been acquainted with the problem area the thesis is situated in, the aim and research questions to be dealt with, the limitations placed on the study and its relevance to library and information science. In the background section of the thesis, OA, IRs and bibliometrics will be introduced, both as concepts and as research fields and from both historical and contemporary perspectives. Included here will be background on bibliometric research evaluation practices in Sweden and internationally. In the theory section, the theoretical framework of the thesis will be specified, with an emphasis on the work of Richard Whitley on the sociology of science and selected theories on citation behavior. Next, the methodology of the study conducted in the thesis will be described, going through source selection choices and the data compilation procedure, as well as the results presentation process. The limitations inherent to quantitative studies of science will also be named. The results and analysis of the study will then be presented, first with a descriptive analysis of LUP's OA content and then with a comparative analysis of LUP's download statistics and WoS's citation analysis. This will be followed by a discussion of the results that ties the descriptive and comparative parts of the analysis together and discusses the results within the study's theoretical framework and in relation to previous research. A concluding section will sum up the findings of the study and mention possible avenues for further research.

2. Background and previous research

This section will provide background to the current study as well as highlight previous research which provides context to the present study. The disposition of this section will be thematic, with first an introduction to OA, followed by an outline of IRs and related research, and lastly a more comprehensive section describing the field of bibliometrics and the role of bibliometrics in research evaluation.

2.1 Open access in brief

Questions relating to OA play a central role in in the present thesis and therefore a clarification on the meaning and use of the term is necessary. The term "open access" was first formulated and defined at the Budapest Open Access Initiative in 2001 (Bailey 2007) as the free availability, online, of scholarly literature where no direct monetary compensation is given to the authors of the works. Besides requiring that the user of the information properly cites the source and respects the integrity of the work, users are free to make use of the work and further copy and distribute it.

Two key factors led to the emergence of the OA movement (Oppenheim 2008, p. 579). Firstly, rapid advancements in information technology in the latter half of the twentieth century allowed for the quick, easy and low-cost dissemination of documents. Secondly, there emerged dissatisfaction from the fact that while taxpayer money often funded research, taxpayers were largely unable to freely access research output. It is important also to note that the rising cost of database licenses had a detrimental effect on the budget of libraries, and indeed libraries have had a major role in the development and promotion of OA alternatives, although as Bailey (2008) discusses, it is not a wholly uncomplicated relationship.

There are two primary branches of OA, the so-called "gold" and "green" paths (Oppenheim 2008, p. 579-581). Green OA, also called "self-archiving", involves the author submitting a copy of a work that has been published elsewhere to an online, freely accessible repository, which can for example be subject- or institution-based (the institution-based repository will be the focus of the present study). In gold OA, the research is published directly in an online, peer-reviewed OA journal. In this case the author, their employer or funder may be required to pay an article processing charge (APC) with submission that is used to support the OA journal. These two routes to OA publishing need not be mutually exclusive; an author can both publish a paper

in an OA journal and deposit a copy to a self-archiving repository. A third possibility for OA publishing is the "hybrid" route in which researchers pay a one-time APC to publish a specific article OA in a journal which is primarily toll-access but has some OA content. This possibility is however complicated by the expensive APCs in high-prestige journals and uncertainty over funding for APCs (Björk and Solomon 2014).

In Sweden, the OA movement has established itself in the research-funding environment. The Swedish Research Council (Vetenskapsrådet), a government agency with a mandate to distribute funding for research in Sweden, has as of January 2010 a policy that research receiving funding from them must publish their results OA (Swedish Research Council 2015). However, this policy only encompasses peer-reviewed journal articles and conference papers, and allows an embargo of 6-12 months depending on the subject field. Books and book chapters are not included in this policy, despite the fact that these document types are an important publishing medium for humanities research. In addition, internationally, scholarly literature publishers are actively opposing such legislation and policies from research funders (Borgman 2007, p. 103, 241).

The possibilities for OA publishing can be different in different academic areas. In a report from Lund University based on information from LUP, Lindh and Wiklund (2010) found that among the humanities and legal science, there is confusion among publishers of both journals and monographs on the legal and financial aspects of OA. However, while journal publishers generally expressed a positive attitude to OA, monograph publishers expressed concerns over financial losses from diminished book sales. The advantages of OA publishing can also vary between academic disciplines. Norris, Oppenheim and Rowland (2008) investigated whether publishing OA led to increased citations, finding that although there appeared to be a citation advantage to publishing OA, this varied substantially between the disciplines included in their study.

2.2 Institutional repositories and related research

The second concept crucial to the thesis is that of the institutional repository (IR). IRs are online databases that collect the scholarly output, in a variety of forms such as journal articles, dissertations and presentations, of a particular institution or institutions (Bailey 2007). These repositories fulfill two main functions for the institution: they provide a means of disseminating the output of the institution and serve as a place to collect information on this output for internal evaluative purposes (Jones 2007, p. 52). In addition, IRs help libraries fulfill their preservation mandate by allowing them to store copies of documents produced by their institution which might otherwise be solely entrusted to a commercial, subscription-based service with no guarantee that the documents will be available in perpetuity (Oppenheim 2008, p. 583). Borgman (2015, p. 102) notes that the interests of authors, libraries and universities are all aligned when it comes to self-archiving.

IRs play an important role in OA. As mentioned above, they are a part of the green OA route in that they provide a place to archive and provide access to

scholarly output. Unfortunately, placing a copy of a work in an IR may be hindered by the work being published elsewhere and the publisher retaining the work's copyright. In some cases this copyright restriction expires after an embargo period or the publisher may allow for a pre-print version of the paper to be archived. In the cases where copyright does not allow for a full-text version of the publication to be made available, bibliographic metadata alone may be placed in the IR, as is the case in Lund University's IR.

IRs have now been in use in various forms for over a decade and they have become the subject of much recent research within the field of library and information science. Many texts deal with IRs on an administrative level, for example Jones (2007) on IR policy and maintenance and Inefuku (2013) on the restructuring of IRs following institutional reorganizations. Several studies have focused on the acceptance of IRs by the academic community, for example, Dutta and Paul (2014) and Stanton and Liew (2012) or Dorner and Revell (2012) on librarians' perceptions of IRs. Other studies have examined the potential of IRs as a complement to traditional evaluation measurements of research impact, for example Bonilla-Calero (2008; 2014). Cho (2014) conducted a keyword analysis on abstracts in Scopus from the growing field of IR research and found that OA was a major topic of discussion within the IR research field, which is in line with the focus of the present study. This thesis therefore is very much at the heart of a relatively new yet growing field of study directed toward investigating and improving IR functionality.

As we have seen in sections 2.1 and 2.2, OA can take a variety of forms and IRs play an important part in many OA publishing channels. As research funding becomes dependent on OA publishing, IRs are likely to be even more instrumental in OA. However, previous investigation has indicated that the conditions for and advantages of publishing OA are dependent on document type and academic subject area, necessitating an investigation into how the IR OA landscape looks in practice, which the present study aims to do. While previous research has investigated IRs and OA, the present study differs from previous work in that it will provide a more comprehensive picture, combining descriptive analysis of the IR with the impact measurements of both citation analysis and download statistics.

2.3 Bibliometrics

The last concept integral to the present study is bibliometrics. This section is split into three sub-sections. First, a historical overview of the development of bibliometric methods will give the reader an understanding of how and why quantitative science studies emerged and how they have been used to describe and evaluate scholarly research output. Second, a description of the to a certain degree interchangeable terminology used to describe the quantitative study of science will help the reader situate the present thesis in the context of ongoing developments in the field of quantitative science studies. Third, a description of the application of bibliometric methodology to the evaluation of research, with a focus on Sweden, the country in which the IR that is the object of analysis in

the current study is located, will highlight the critical evaluation and application of quantitative science methodologies on research performance evaluation.

2.3.1 The evolution of bibliometrics

Though the quantitative study of scholarly literature owes its advent to a long lineage of scientific and technical achievements, from the development of statistical methods to the Gutenberg press which allowed for the mass production of texts without which the large-scale dissemination of books would have never been possible, De Bellis points to the 1917 study by Cole and Eales as being often credited as one of the first true studies utilizing bibliometric methods (Cole and Eales 1917 in De Bellis 2009, p. 6-7). In the study, Cole and Eales mapped, with quantitative methods, progress in the field of comparative anatomy. Their study was a descriptive analysis but also had an evaluative dimension, foreshadowing the future of bibliometric studies. De Bellis (2009) discusses three assumptions present in Cole and Eales' study on which bibliometrics rests its foundations: the final, stable object of measurement that is the published text; the common goal of describing and evaluating science; and the inherent weakness of quantitative analysis based in ultimately arbitrary decisions made in the compilation of the data. Another assumption is that of the ultimately cumulative nature of science: that the achievements of individuals can be attributed to the past labor of many others, on a single and inescapable trajectory toward universal truths (De Bellis 2009, p. 10-11).

Bibliometrics owes much of its foundations to the evolution of statistical measurements and early studies of scientific communication. The first half of the twentieth century saw the pioneering work by Cole and Eales mentioned above but also for example Gross and Gross' (1927) study on citations in the field of chemistry and Gosnell (1944) on the life cycle of books in college libraries. Other important forerunners who contributed to the theoretical mathematical framework of bibliometrics were Lotka (1926), Bradford (1985) and Zipf (1965): the sources of the so-called bibliometric laws that lay the foundation for future mathematical studies of scientific literature. These laws are in no way exact reflections of the way scientific communication can be expected to behave, but rather the laws and the work done subsequent to them provide a framework for the interpretation of bibliometric analysis (De Bellis 2009, p. 75).

Another important contributor to bibliometric theory, in specific, theories on the growth of science, was Derek J. De Solla Price, whose book *Little Science, Big Science* is a landmark in the field (Price 1965). His work rests on the basis that science is a cumulative enterprise whose ultimate manifestation is the scientific publication, and is thus amenable to quantitative measurement in the pursuit of knowledge about the evolution of science. From his studies, Price came to the conclusion that science has had an exponential growth and will inevitably reach a saturation point. Furthermore, because science was in a constant state of exponential growth, most of the science ever produced will be fairly current (ibid, p. 11). Price also discussed the tendency for citations to decrease after publication, with differences in citation trends over time and

between subject fields (ibid, p. 77-82). Price's *Little Science, Big Science* and other important early works non-withstanding, it was the rapid advancement of information technology in the latter half of the twentieth century that was the true springboard for the development of bibliometric methods in the study of scientific literature, both within the research field of bibliometrics and in its practical application in research evaluation.

No account of bibliometrics can be complete without an account of the advent and impact of the citation index. The first citation index, the Science Citation Index, the precursor to Web of Science, was developed by information science pioneer Eugene Garfield in the 1960's (De Bellis 2009, p. 32-39). The Science Citation Index was originally developed as an information retrieval tool at a time when Cold War investments in science were driving a need for more efficient organization of scientific information. Garfield had ruminated on the various possibilities for a science citation index in his 1955 article *Citation Indexes for Science* (Garfield 1955), but his ideas were mainly focused on its ability to allow scientists to more easily trace the transfer of ideas (though he does mention very briefly the possibility of assessing an article's "influence" through citation tracing (p. 111)). The potential for such an index in analyzing networks of scientific publications for the mapping and assessment of scientific production was an exciting potential secondary application in the aftermath of its creation (De Bellis 2009, p. 32-39). With time, the citation index became an indispensable tool for the emerging field of the quantitative evaluation of scholarly impact, that is to say bibliometrics/scientometrics, fostering a symbiotic relationship between the field of citation analysis and the index that enabled it. The dawn of the Internet age promoted the citation index's move to the Internet. Today, the citation index, now called Web of Science (WoS), is owned by Canadian information company Thomson Reuters. A citation index like the WoS allows for complex types of searching that escape the abilities of traditional databases (De Bellis 2009, p. 42). Citation indexing services are now also provided by Google Scholar and Scopus, with substantial variation between databases (De Bellis 2014, p. 33; see also discussion below, section 4.1.2).

The development of bibliometric methods has since come into its own as a field of research. It contains all the hallmarks of a mature research field, with dedicated academic journals, for instance *Scientometrics*, and miscellaneous conferences, reviews, societies, prizes etc. (De Bellis 2009, p. 15-17).

2.3.2 A quantity of terms for the quantitative study of science

Many different terms have been used to describe the quantitative study of scientific communication, for example scientometrics, informetrics and bibliometrics. De Bellis (2009, p. 2-5) discusses the meanings of these different terms and how they are to a certain degree interchangeable, in that they all deal with the quantitative study of communication. According to De Bellis, where they differ is in their emphases and their areas of interest. Bibliometrics focuses on the material expression of recorded information, whereas informetrics takes on a broader scope in that it encompasses all quantitative measurements of communication. Informetrics is in this sense an umbrella term for all metrics.

Scientometrics, on the other hand, limits itself to the study of a certain type of communication, that within the scientific establishment. Terms like webometrics, netometrics and cybermerics specify Internet communications and web resources and the unique aspects of Internet transactions.

This thesis then falls under many of the terms used to describe quantitative studies of science by combining traditional bibliometric practices of citation analysis with usage data from downloads. The term bibliometrics is applicable to the study's methodology due to the fact that publication objects' bibliometric metadata will be analyzed; informetrics as the research publications are a kind of communication in the broader sense included in informetrics as well; and scientometrics because the communications being studied are scientific in their content, purpose and origin. Since the study conducted in this thesis utilizes data derived from download statistics, the study falls under webometrics as well. For the purposes of this thesis, the terms bibliometrics and scientometrics will be used to refer to the methods and analysis used in the study. This is a pragmatic choice based on these terms being widely used and accepted for describing quantitative studies of science.

Traditional bibliometric/scientometric studies and practices have focused on document-to-document relationships via citation analysis as a measure of a document's impact on the scientific community. However, this tradition is being challenged by the emergence of new methodological areas for the quantitative measurement of research impact, reflected in emerging sub-fields within quantitative studies of science. An example is the recent development of "altmetrics". In Priem, Taraborelli, Groth and Neylon's (2010) manifesto for the usage and development of altmetric indicators, altmetrics is described as the use of data derived from social media communications to measure science impact. They base the need for altmetric indicators on the failure of traditional indicators to keep up with the fast-evolving nature of scientific communications. They aim specific criticism at the cumbersomeness of citation analysis and insist that modern measures of research impact must be leveled at the individual article and follow the impact of science outside the realm of academia. In short, altmetrics is about harnessing the potential of alternative sources of data in the measurement of research impact, for example through sources such as social media, reference managers, or collaborative encyclopedias (Priem, Groth and Taraborelli 2012). The present thesis does not fall under this definition of altmetrics, as it does not derive data from social or collaborative media. However it is closely related in that it attempts to investigate alternatives to citation analysis in quantitative measurements of science.

Altmetrics has garnered much attention from bibliometricians. Haustein et al. (2013) looked at the potential of altmetric indicators by assessing the coverage of bibliometricians' own research output in social media and bibliometricians' attitudes towards the use of altmetrics in research evaluation, finding that 71.8% of surveyed bibliometricians thought that downloads or views of articles had potential as an impact measurement (2013, p. 1159). Bornmann (2014) investigated the definition, benefits and drawbacks of altmetrics measurements

from a scientometric perspective. Priem (2014) provides a good overview of this emerging field and points to the need for further research.

Another emerging area in quantitative studies of sciences is usage bibliometrics. Kurtz and Bollen (2011) identify a number of disadvantages with the current dominance of citation-based analysis in bibliometrics, including the primacy of journal articles and publication delays (the delay from the publishing of the cited publication to the publishing of the citing publication). They point instead to the potential of usage bibliometrics, or the use of data generated from the access of electronic resources. Usage data have the distinct advantage of encompassing all web-based resources, being available more immediate to publication and potentially offering a broader view of the scholarly information consumption process. Usage data can also work in conjunction with traditional bibliometric indicators. However, unlike citation data, the meaning of usage data has only recently come under the scrutiny of the bibliometric field and therefore this type of data must be used with caution. For example, the user's intention when downloading an item cannot be assumed (Kurtz and Bollen 2011; see also Haustein 2014, p. 329-334).

The present study, in its use of download statistics from an IR, investigates the possibilities in science measuring that download statistics provides, and is thus firmly situated within usage bibliometrics. As usage data is thus far an under-investigated area of quantitative science studies, the present study will be cautious in drawing conclusions from the analysis of download statistics. As discussed by Kurtz and Bollen, the relationship between usage data and citation analysis is a complicated one (2011, p. 19-27). The types of documents involved, their users/citers, and the age of the document all play a role in the outcome of the analysis. Furthermore, an article may be used more *because* it has been cited. In addition, citations are the measurement of impact on a specific audience – fellow scholars – while download measurements potentially capture the impact on a wider audience (ibid, p. 27). The present study avoids these complications by analyzing the relationship between citations and downloads where specific parameters have been met: the publications are all quality controlled, scholarly articles in the IR and indexed in WoS. The discussion of the analysis and results will be sensitive to possible audience differences between faculties, with regards to both citations and downloads. Indeed, the possibility that downloads offer a look at impact on a broader audience, thus adding a novel measurement compared with what traditional citation analysis offers, will be viewed positively.

From the plethora of terminology and sub-fields it can be inferred that there is a need for an abundance of perspectives in the quantitative study of science and that the field is fertile for ongoing research. In particular, studies are needed that explore how usage data derived from electronic resources can be used alone and in combination with established methodologies in bibliometric studies, which the current thesis hopes to in part provide.

2.3.3 Bibliometric evaluations of science in Sweden and internationally

The era of big science had its inception in post-World War II investments in science and research (De Bellis 2009, p. 11). Not only did this lead to a sharp increase in the quantity of scholarly literature and a need to find systems that could handle this new information wealth, but also, eventually, the necessity of a means of evaluating scientific production, although this trend was tempered at first by a predominant view of science as ultimately progressive and self-regulating (De Bellis 2009, p.11-14). It was in this climate that Garfield's citation index, the genesis of which is discussed above, found fertile ground for its inception. The field of evaluative bibliometrics, with citation analysis at its core, was born.

The IR analyzed in this report is situated in a university in Sweden, where bibliometrics has a firm foothold in evaluations of scientific production. In Sweden, the use of bibliometrics as a tool for research evaluation began in the late 1960's with the Natural Science Research Council ordering citation analysis of broad disciplinary fields for the purposes of surveying natural science research areas and assessing their own funding policies (Engwall & Nybom 2008, p. 43). These evaluations, however, had little bearing on actual implemented policy (ibid). This situation has since changed, and from 2009, performance-based resource allocation partially based on citation and publication analysis comprises a substantial portion, currently 20%, of the direct government funding to higher education (Swedish Research Council 2014, p. 14).

The Swedish Research Council has recently come out with a report on resource allocation to Swedish higher education institutions based on quality and performance (Swedish Research Council 2014). Citation analysis is included in their suggested source allocation model as an assessment tool in the scientific/artistic quality category (ibid, p. 11, 31). Furthermore, the proposal included a development of the national research publications database, SwePub, which gathers data on research publications from higher education institutions' IRs (ibid, p. 59; National Library of Sweden 2014). Well-functioning IRs will be an important precondition for the performance-based research evaluation model proposed by the Swedish Research Council. However, the model did not include any preferential treatment to OA publications, though the report acknowledges that this might present a conflict between the suggested model and a national policy for OA (ibid, p. 19). However, the report did include a category for research impact outside academia, which is an indication that a wide impact of research output is desirable, something that the OA movement also tries to promote by making research freely available (Swedish Research Council 2014, p. 11).

Hammarfelt, Nelhans, Eklund and Åström (2015) investigated the use of bibliometric indicators for resource allocation at Swedish institutions. They found that a majority of Swedish higher education institutions employed some form of bibliometric resource allocation system (ibid, p. 10), though these systems varied greatly from institution to institution based on factors such as

size and research profile. The authors connected the implementation of performance-based resource funding systems with a culture of evaluation, stating, "The idea of evaluation is ingrained in the academic system and few question the need for, and overarching goals of, the evaluation" (ibid, p. 20). They also found that local publication databases – or IRs – played a substantial role in the bibliometric evaluation processes. As the present study utilizes publication data from an IR, the results have the potential to highlight the role of IRs in research evaluation.

On the international front, performance-based research funding systems exist on the national level in many countries (Hicks 2012). These systems are both heterogeneous and complex, varying in method from peer-review to metric indicators, or a combination of the two (ibid). As the current proposal from the Swedish Research Council testifies, a scarce five years after the implementation of a new model for allocation of performance-based research funding, further improvements are necessary in indicators of research impact for use in the evaluation of science. The use of bibliometric measurements for evaluating performance has ethical aspects as well that are not fully understood (Furner 2014), making evaluative bibliometrics a turbulent territory.

One of the prominent issues in evaluative bibliometrics is subject field differences, as traditionally, the natural and life sciences have been the primary sources of data in bibliometric investigations. Recognizing that the differences in publication patterns in the humanities necessitated a different approach from bibliometric studies than the traditional indicators developed with the natural and life sciences in mind, Hammarfelt (2012) explored the applicability of bibliometric methodologies to the research field of literary studies. His purpose was threefold: to investigate the social and intellectual organization of the field of literary studies and how this potentially affected citation patterns; to discover how bibliometric methods could be made more applicative to the study of the research field of literary studies; and to explore the effects of bibliometric evaluations of research in the humanities in general. He came to the conclusion that bibliometric methodologies must be modified to include non-English and non-journal publications and take into consideration differences in citation patterns in order to fairly evaluate humanities research. Alternatively, an altogether different bibliometric approach may be considered, such as usage data gathered, for example, from an IR. From this perspective, the present study and other similar studies take on an important role in the development of bibliometric indicators more appropriate for the humanities.

Hicks and collueges (2015) recently came out with the Leiden Manifesto for research metrics. This manifesto contains a ten-point guideline on good application and interpretation of scientometric indicators in research evaluation, and is a reaction toward the increasing integration of scientometric indicators in research evaluation. These guidelines include such points as using quantitative measurement to support qualitative assessment, protecting research of local interest in the evaluation process, accounting for subject field differences in publication and citation, and regularly assessing the validity of the scientometric indicators used in research evaluation. These guidelines can

seem common sense but are often overlooked in the increasing routineness of research evaluation reliant on metrics. Though the current study is not in itself geared toward research evaluation, it is situated in a climate of the proliferation of quantitative evaluation practices in science, and recognizes the need for caution and consideration in the application of scientometric methodologies. In this sense, the present study hopes to add the discussion of the applicability of scientometric indicators in science evaluation, especially in the case of the interpretation of download statistics as an indicator of research impact.

3. Theoretical framework

The theoretical framework for the thesis will be based on Robert K. Merton's and Richard Whitley's works on the sociology of science. Whitley's theory in particular will inform the interpretation of the results in relation to IRs, OA, and the evaluation of the sciences. Whitley's theories have been successfully applied in bibliometric studies previously (for example Hammarfelt's (2012) work within bibliometrics described above and Fry and Talja's (2007) study on scholarly communication described below), providing proof of the theory's applicability to the topics of concern in the present thesis. These and other works presented below will also provide further concepts and models for the interpretation of the results gathered in the present study. Additionally, section 3.3 will provide key concepts for citation analysis theory which will be used as tools for the interpretation of the citation analysis conducted in the study.

3.1 The organization of the sciences

The overriding theory on science and scientific production which is the point of departure for this thesis is Whitley's sociological approach as outlined in his book *The Intellectual and Social Organization of the Sciences* (2000). Whitley understood scientific output as the result of the social and intellectual organization of respective scientific subject fields and that science itself was unique from other enterprises in its constant drive for novelty. These scientific fields were thusly subject to changes within and around the work organization. According to Whitley, "Fields organized and controlled in different ways produce differently organized knowledge…" (ibid, p. 33-34). Additionally, reputation as acquired by the perceived relevancy and usefulness of a researcher's work to their colleagues is seen as a unique and integral aspect in modern science and makes necessary a robust, formal and public communications network between scientists. Scientific reputations in turn can lead to material rewards such as funding and other resources for conducting further research.

Furthermore, Whitley introduces a number of important contextual factors to the intellectual and social organization of the sciences, one of which is audience structure (2000, p. 234-238). Audience structure, or the diversity of audiences to which researchers address their results and which can affect their reputation, varies substantially between subject fields. A uniform audience for a scientific field encourages consensus on research goals and procedures, whereas when audience diversity is high, the reverse is true. Audiences can exert varying

degrees of reputational control as well, that is to say, not all audiences are equal.

To outline a framework for understanding the organization of the sciences and the influence of changes on that organization, Whitley (2000) introduced two key concepts: "mutual dependence" and "task uncertainty". Mutual dependence describes the extent to which scientists are reliant on their colleagues to make substantial contributions and set collective goals as well as a field's dependence on other scientific fields (ibid, p. 112-113). Mutual dependence is further divided into two aspects: functional dependence, or "the extent to which researchers have to use the specific results, ideas, and procedures of fellow specialists in order to construct knowledge claims which are regarded as competent and useful contributions" and strategic dependence, or "the extent to which researchers have to persuade colleagues of the significance and importance of their problem and approach to obtain a high reputation from them" (ibid, p. 88).

The task uncertainty dimension, on the other hand, defines the degree to which research outcomes and strategies are predictable, visible and stable within and surrounding a research field (Whitley 2000, p. 148-149). Task uncertainty is further divided into two aspects: technical task uncertainty, or "the extent to which work techniques are well understood and produce reliable results" (ibid, p. 121) and strategic task uncertainty, or "uncertainty about intellectual priorities, the significance of research topics and preferred ways of tackling them, the likely reputational pay-off of different research strategies, and the relevance of task outcomes for collective intellectual goals" (ibid, p.123). The consequences of these intellectual and social aspects of functional and strategic dependence and technical and strategic task uncertainty affect the work organization in scientific fields.

These terms can be applied in understanding the connection between IRs, OA, research evaluation and work organization in the sciences. According to Whitley, changes in mutual dependence can affect both the control over knowledge production and dissemination as well as potential audiences (2000, p. 113). This can have consequences for researchers' perception of OA and the role of IRs in their research field, and whether they see a tangible benefit in increasing the audience for their research. On the task uncertainty side, Whitley maintains that task uncertainty is higher in fields where resource allocation in a field is funded through diverse channels and evaluation standards are more flexible (ibid, p. 148).

Whitley expands on his theoretical framework directly in relation to the evaluation of science (Whitley, 2008). Building on his previous work, Whitley examined the consequences of research evaluation systems. He differentiates between "strong" and "weak" evaluation systems by the degree of standardization of evaluation procedures (ibid, p. 9). Whitley (2008) outlines five consequences of strong evaluations: stronger coordination in research fields; less uncertainty in collective strategies and goals within a field; decline in diversity of approaches to research; inhibition in the development of new

fields of enquiry; and the consolidation of resources and skills in elite universities. He then relates these factors to differing funding regimes within academics at the national/university level and to differences in the organization of scientific fields. As discussed above, Sweden has a funding scheme partially allotted through performance evaluations based on, among other factors, citation and publication analysis (Swedish Research Council 2014, p. 14-15), and this will have consequences on the way work is organized.

Hammarfelt and de Rijcke (2014) discuss research evaluation in Sweden from the perspective of Whitley's (2008) strong/weak evaluation dichotomy, asserting that Sweden's performance-based research evaluation system is strong. They used a mixed methodology of interviews and – similar to the study presented in this thesis – publication data from an IR analyzed by publication type to investigate how the publication behavior of researchers at a humanities-oriented faculty at Uppsala University changed in response to the implementation of a performance-based evaluation system. Their investigation showed an increase in journal article publishing, especially English-language journals. Additionally, if download statistics were to be used as a strong evaluation measure of research impact, this would be expected to have consequences for the scientific fields based on their current organization and funding structures.

In light of the theoretical framework provided by Whitley, IRs, IR content and OA publishing behavior can be understood within the context of a larger knowledge-producing apparatus, within the work organization of the university and faculties. The IR, containing the collected scientific output of the university and faculties, reflects the work organization and production of the scientific fields at work in the university. Changes in communication methods and evaluation markers affecting the work organization of the fields, for instance the OA movement and changes in the evaluation of the sciences, will thus also be reflected in the IR. Together, IRs, OA, and changes in the evaluation of research constitute substantial potential drivers of change in the organization of the sciences. The results of the study in the present thesis will be analyzed within Whitley's theoretical understanding of the sciences as work organizations producing knowledge in a context of social interactions, and in the context of previous research as presented earlier in the thesis, in an attempt to interpret how the IR LUP reflects current scientific work organizations. Further, the potential role of download statistics derived from the IR as a measure of research impact will be discussed with a point of departure in the results of the study and Whitley's theories on the organization of the sciences and its relationship to research evaluation. The theoretical framework will help in analyzing the results of the study with an understanding of the complex interaction between research activities, the OA movement and the evaluation of research.

3.2 Scholarly communication and subject fields

Fry and Talja (2007) identify the need to create a theoretical framework for understanding how the varied intellectual and social structures between

academic fields affected the production and use of digital scholarly resources. They identified this need in relation to the investments being made in the development and maintenance of digital information repositories and communications for scientific information exchange. To this end, they applied Whitley's theoretical concepts of mutual dependence and task uncertainty, which they stress are relative concepts made meaningful in comparisons between the objects studied. Having first conducted in-depth interviews with representatives from seven academic fields in order to establish the degree of mutual dependence and task uncertainty in the fields, they proceeded to analyze scholarly communication from the chosen fields in the form of scholarly mailing lists, academic homepages, and what the authors termed "scholar-produced decentralized digital resources", which they defined as characterized by being organized and made accessible by networked groups of researchers and include such diverse resources as bibliographies, e-journals and non-proprietary software made freely available on the Internet. Fry and Talja's study came to the conclusion that an understanding of field-specific mutual dependence, task uncertainty and target audience must be achieved in order to interpret scholarly communication practices and design scholarly communication systems. According to Fry and Talja, these field-specific differences can explain researchers' unwillingness to adapt to promoted communication models designed and tailored to a field with different mutual dependence and task uncertainty characteristics.

Fry and Talja's (2007) study is relevant to the current thesis in that it provides a model for the application of Whitley's theory and concepts specifically in the analysis of differences between subject fields. They show that work organization as interpreted through Whitley's concepts of mutual dependence and task uncertainty can be utilized in order to explain field differences. Similarly to Fry and Talja, the current thesis will examine the research production in the IR through the lens of Whitley's concepts. The theoretical model will then provide a basis for understanding potential differences in communication behavior identified by the IR. However, due to the methodological differences between Fry and Talja's study and the present study, the depth of analysis achieved by Fry and Talja in the application of Whitley's theory cannot be attained in the present study. This is due both to the present study being limited to analysis on the faculty level as compared to Fry and Talja's more fine-grained subject analysis and the qualitative nature of Fry and Talja's interview methodology compared to the present study's quantitative bibliometric approach. However, the study will aim at using Whitley's theories on a more general level to explain discrepancies in research output between faculties, especially in relation to OA in the IR studied. Nonetheless, Fry and Talja's results will inform the analysis and discussion of the present study's results.

3.3 Citation theory

Citation indexes provided a novel way to create order in scholarly literature for information retrieval, in which the relevance of a document is judged by its perceived relevance by previous authors by way of the bibliographic reference

(De Bellis 2009, p. 31). As mentioned previously, evaluative bibliometrics has relied heavily on citation indexes and citation analysis as a means of assessing research impact; the present study will also be partially comprised of analysis built on the analysis of citation data. It is important, then, to examine the phenomenon of the bibliographic citation and its place in scientific communication. Many previous theorists have examined bibliographic citations' function in scholarly communications and the consequences of this for citation analysis; some key and interesting theories will be mentioned here.

As mentioned earlier, science is a socially situated undertaking; scientists collaborate and communicate with a network of other scientists. Robert K. Merton, an early contributor to the sociology of science, recognized that scientists are bound by norms of conduct. The foundations of these norms revolve around the nature of the scientific pursuit as rational, skeptical and altruistic in its ambition to advance scientific knowledge without regard for personal gain (Merton 1973, p. 268-278).

Interesting to the present study is Merton's norm of "communism": the idea that scientists contribute their work to the body of science without any material rewards, save the right to be mentioned in conjunction with other scientists' use of that contribution (Merton 1973, p. 273-277). Since this recognition is the sole reward for the work put in by the scientist, and contributions must be original, there exists a strong incentive for scientists to lay claim on their contributions and seek others to acknowledge this claim (ibid). This exchange of recognition between scientists can take the form of the bibliographic citation (Merton, 1988). Merton describes it thusly:

> … the institutionalized practice of citations and references in the sphere of learning is not a trivial matter. While many a general reader – that is, the lay reader located outside the domain of science and scholarship – may regard the lowly footnote or the remote endnote or the bibliographic parenthesis as a dispensable nuisance, it can be argued that these are in truth central to the incentive system and an underlying sense of distributive justice that do much to energize the advancement of knowledge.
>
> *Merton, 1988, p. 621*

Merton also recognized what he called "the Matthew effect": the tendency for scientists who have previously been acknowledged for their achievements to continue to accrue praise, regardless of the quality of their continued achievements (Merton 1968). This concentrates academic preeminence into an elite few scientists, disproportionately credited for their work compared to their actual contributions. Merton's Matthew effect (ibid) has practical consequences for understanding citations, particularly how citations pattern in a citation index. The Matthew effect leads to scientists preferentially citing the works of elite, prestigious scientists to the disadvantage of the potentially better quality and more relevant works by less visible scientists (Merton 1988). Though Merton worked closely with Garfield and was positive toward citation indexing, he was also an early voice of caution in the use of citation analysis. In the foreword to Garfield's *Citation indexing: its theory and application in science, technology, and humanities* Merton puts forth the idea of "obliteration

by incorporation", whereby texts have become so integral to a field that they are taken for granted as part of the common knowledge and cease to be directly cited (Garfield 1979, p. ix). Consequently, these essential works will cease to accrue proper credit in the form of continued citations (for a recent review of the phenomenon, see McCain (2014)). As the present study will consider only relatively contemporary scientific literature's amassed citations, the "obliteration by incorporation" phenomenon should pose no risk to the interpretation of the works' impact on the field; however, it is important to remember that citations may not reflect the total impact of a scholarly work on its field.

Small addressed the issue of how to interpret the bibliographic citation from the standpoint that the citation functions as a symbol for the ideas contained in the cited work (Small 1978). This viewpoint of citation as symbol relates back to Garfield's reasoning that citations are representations of the transfer of ideas (Garfield 1955, p. 109-110). The advantage of this point of view is that it partially bypasses the common critique of citation analysis that there exists multiple reasons for why citations are given in academic texts by seeing the citation as a symbol whose representational value has the potential to become fixed by its repeated use in a community of peers (Small 1978, p. 377).

In contrast to Merton and Small's citation theories is the social-constructivist approach to citation of Latour and Woolgar (1979) and Gilbert (1977). Citations here are understood as a rhetorical strategy on the part of the author to bolster the status of their own work, for example by intentionally citing prestigious works or authors, rather than to acknowledge intellectual debt. However, more recent empirical studies refute this hypothesis, instead favoring Merton's normative citation theory (Baldi 1998; White 2004).

The citation theory thus far has taken for granted that there is a generally unified behavioral pattern in citation; this is an oversimplification given that citations structure differently in different subject fields. In reviewing the literature on the characteristics of citation patterns in the humanities, Hammarfelt states:

> Studies of citation characteristics in the humanities show that the type of publication that is most frequently cited is the monograph, the age span of cited sources is broad, the rate of obsolescence is low, languages other than English play an important role, and self-citations are rare.
>
> *Hammarfelt, 2012, p 34*

It is important to recognize that citation patterns are connected to other factors in the organization of the discipline, factors that can be related back to Whitley's theory on the social and intellectual organization of the sciences.

Building on Whitley's work, Åström and Sándor (2009) put forth a model for understanding patterns of scholarly communication and citation analysis which takes into account that scholarly communication is not always primarily cumulative in nature, as assumed in most studies of science based on citation

analysis. Instead different fields can exhibit primarily cumulative, negotiating or distinctive citation styles. This provides a more nuanced picture of how and why authors cite, which has consequences for the interpretation of the results of citation analysis. This further highlights the need for alternative measures of research impact to provide a more nuanced picture of research impact, which download statistics from IRs, as examined in the present study, have the potential to provide.

This thesis makes no attempts to answer the *why* of how authors cite. Following the tradition set by early bibliometric theorists, the thesis will analyze the data under the assumption that the bibliographic citation is a marker of an author's intention to acknowledge the intellectual or methodological contributions of a previous author and therefore can be used as an indicator of an individual document's impact. At the same time, it is recognized that the citation patterns can be interpreted in a multitude of ways which can have consequences for how the role and meaning of the bibliographic citation is interpreted. The current study will therefore relate back to the citation theory and concepts presented in this section when analyzing the results of the citation analysis.

4. Method

The methodology of the present study can alternately be termed bibliometric or scientometric (see discussion in section 2.3.2). The study is a descriptive bibliometric/scientometric investigation into the content of LUP in relation to its OA content. It is important to note that the present study examines OA in relation to IR content – the publications examined in the study may be available OA from another repository or OA journal homepage (Lund University Libraries n.d.). The study also conducts and compares traditional measurements of research impact in the form of citation analysis with less established, alternative measures in the form of download counts from the IR LUP. The section below will go through the source data selection process, the data collection and analysis procedures, and discuss methodological considerations for the study.

4.1 Source data

Much consideration must be given to which databases are chosen for a bibliometric study; as previously mentioned, database construction procedures can have consequences for the results of a bibliometric analysis. Neuhaus and Daniel (2008) identify six factors to consider when choosing which databases to use for a citation analysis: coverage, accuracy of data, data fields, browsing options, search options and analytical tools. For example, to carry out a citation analysis, the list of cited works for all publications in the database must be indexed. The present study makes use of both data retrieved from Lund University's IR and the WoS citation index, which each have their own characteristics that must be taken into account.

As mentioned previously, the scope of the study is limited to three faculties at Lund University: the Faculty of Medicine (FM), the Faculty of Engineering (FE) and the Faculties of Humanities and Theology (FHT). These faculties were chosen because of the divergent characteristics of their subject fields, allowing for an analysis and comparison of the results in subject fields where the socio-scientific behavior has historically been divergent. Faculties were chosen instead of other categorizations of academic subjects/disciplines/fields partially because of the problems inherent in establishing these types of categories and partially in acknowledgement of the importance of faculty organization in the academic context, especially in the knowledge organization of the IR. A further narrowing of the scope limits the data collection to a span of five years from 2008 to 2012; this provides enough data while taking into

account the typical delay between publication and citation that would skew the data analysis if years more contemporary to the study examined.

4.1.1 Lund University Publications and download counts

Lund University Publishing, or LUP, is an IR hosted and maintained by Lund University Libraries. Bibliographic metadata on all research output from the university is collected and stored in the database; full coverage has been maintained for research output at Lund University since 2008. If the text is not published elsewhere or in cases where copyright allows, a full text version will be made available for download directly from the database. The database contains approximately 130,000 records. Approximately 87% of the records contain metadata only, 12% contain a full-text file available for retrieval, and 1% are available only for Lund University students and staff. Bachelors' and one- and two-year masters' theses are not covered by the database and are instead stored in a separate database, LUP Student Papers. For research published in WoS-indexed publications there exists an automatic update system in LUP. Once a month, the database imports the bibliographic information of new items registered in WoS associated with Lund University (Faculty of Medicine 2014b). Faculty librarians then make sure that the data is correct and add additional information as needed (ibid). Publications in LUP become searchable in Google Scholar, making the IR a valuable tool in spreading the university's research and further increasing the importance of full coverage and adequate quality for the metadata stored in the IR.

The LUP database records contain metadata in a wide range of fields, including author, title and year but also for example institution, document type and "quality controlled" status. Advanced searching is available by specific metadata fields and post-search filtering is also enabled. The results of searches can be exported to an Excel file. LUP also contains a number of tools for analysis of the download statistics of the full-text documents, called LUP Statistics. The searcher can view download statistics for individual papers over time and additional information on which countries the downloads have taken place from. Furthermore, statistics are available for the whole university and by faculty for downloads over time, the most downloaded document type, most downloaded author, most downloaded dissertation, and most downloaded other document type. The results are shown in tables, graphs and pie charts where appropriate. To facilitate the study conducted in this thesis, the administrators of LUP provided the author of this paper with a URL address to a complete list of all LUP document id numbers tied to documents and those documents' download statistics. The list is updated daily.

The LUP database has several limitations that can have consequences for the study. As the information in LUP is provided by researchers and manually controlled by faculty librarians, there is a good quality to the bibliographic information, though there can exist consistency issues. However, as Hammarfelt et al. (2015, p. 17), noted in their study on institutional use of bibliometric indicators for resource allocation at Swedish universities, when researchers submit records themselves, records are subject to the researcher's own judgment on whether a publication has been peer-reviewed or not,

especially since the concept of what constitutes a peer-reviewed publication can vary across fields. Here, there may be a temptation for researchers to attempt to skew bibliometric data in their favor by listing non-peer-reviewed items as peer-reviewed. In addition, there are differing routines in the data collection process between the faculties at Lund University, with some faculties, for example at FM, actively checking the copyright status of new research publications produced by FM researchers and actively requesting the authors' permission to upload the item in LUP (Faculty of Medicine 2014a).

Another factor that must be taken into consideration is that researchers may choose to make their research available OA through other platforms, for example in gold OA journal websites and subject-based repositories, instead of in LUP, meaning that LUP does not represent the totality of OA activity by Lund University researchers through all OA channels. However, as the thesis focus is on precisely the role of IRs, this is not a hindrance for the study. Additionally, when it comes to download statistics, data may be artificially inflated due to for example multiple downloads by the same persons or by search software, though according to the LUP website, "Efforts have been made to exclude downloads by robots and track irregular download activities" (LUP Statistics n.d.).

4.1.2 Web of Science

Web of Science (WoS) is a multidisciplinary citation index, or rather a collection of indexes that can be searched together, maintained by Thomson Reuters. As explained above, the index not only indexes traditional bibliographic metadata such as author and title, but also indexes reference lists. Coverage of scientific journals in WoS is not all-inclusive; from its inception, WoS has strived to include those top journals that represented the core of scientific literature in their respective areas (De Bellis 2009, p. 39-41). WoS has been developed with scientometrics in mind, offering tools directly in the search interface for citation analysis, as well as enabling downloads of up to 500 posts at a time.

As mentioned above, WoS has in recent years received competition from citation index services from Scopus and Google Scholar and the decision to use WoS in the study was both a conscious and practical one. WoS is a database that has been used in citation analysis for many years, making it a well-understood database among practitioners of bibliometrics and making the results of the study amenable to more direct comparison with previous and future studies. In addition, and perhaps most importantly, the WoS reference number was included in the bibliographic metadata from LUP, providing a search marker for retrieving LUP publication items' posts from WoS. In this way the WoS citation index was a pragmatic and obvious choice. WoS is also the database used in the Swedish Research Council's citation analysis (Swedish Research Council 2014 , p. 33-34), making the findings of the study more directly relevant to the discussion on research evaluation practices in Sweden.

It is important to mention that the three citation index databases WoS, Scopus and Google Scholar of course differ in their methods and coverage, and studies

have shown that these differences can result in differing bibliometric analysis results (for example, see Delgado & Repiso (2013); Bharathi (2013); Bar-Ilan (2010, 2008); Bakkalbasi, Bauer, Glover & Wang (2006); Meho & Yang (2007)). For example, Delgado and Repiso (2013) and Meho and Yang (2007) both noted a wider coverage of non-English language journals in Google Scholar. Another important aspect of the WoS citation data is that it includes by default self-citations in the citation count, that is to say, when an author cites their own previous work or when a journal or institution encourages authors to cite their own institution. Some object to the inclusion of self-citations in bibliometric indicators, whether it is authors citing themselves or institutions or journals encouraging the citing of their own publications/institutions, claiming that this artificially inflates citation impact. However, according to De Bellis (2009, p. 184-185), scientists may legitimately wish to cite their own work when it is a valuable contribution to the research at hand, and that in any case, self-citations' inflating influence is mitigated when sufficient amounts of data are analyzed.

4.2 Data compilation

The source data for the study was downloaded from WoS and LUP over a period of two weeks in February 2015. The searches were first made in LUP for publications 2008-2012 for the FM, FE and FHT, limited to those indexed as "quality controlled", as the study will focus only on the scholarly output in the IR. The term "quality controlled" is used in LUP in a similar way to "peer review", but more broadly and subjectively applied; however, this excluded dissertations and monographs, as they were not indexed as "quality controlled". This may limit the dataset somewhat when looking at OA publication by type, but should however not be of much consequence to the correlation analysis of download statistics and citations as monographs have low coverage in WoS. It would therefore likely not have added substantially to the dataset requiring both LUP OA status and indexing in WoS.

For WoS, the database Core Collection was utilized for the searches because of the ability to search by "Accession Number", which, in the cases where an item of research output from Lund University is available, is indexed in LUP as "ISI" or "WOS" number in the field "externalidentifier". Because of limits on the number of posts that could be downloaded from the LUP and WoS (1500 for LUP and 500 for WoS), the searches were limited by year and department for LUP and downloaded in batches of up to 500 in WoS. In the few cases where only a PubMed number was indexed in LUP's externalidentifier field, a separate search was made by PubMed number in WoS in order that the dataset be as comprehensive as possible. The data from LUP and WoS was collected in an Excel document and the WoS or PubMed accession number used to correlate citation frequencies to download statistics. For the purposes of the study, any publication that was possible to download in full text from the database was counted as OA.

Due to the some of the searches having to be conducted on the department level, some duplicates were generated in the dataset for the faculties. These

duplicates were removed. However, duplicates between faculties were kept – this is because the present study is interested in the totality of the research output indexed in WoS for the three faculties. 190 duplicates were present in the dataset in total out of 17924 posts, that is to say 190 posts were associated with more than one faculty. These publications are the products of cross-faculty cooperation. An example of this is a number of publications shared between FM and FHT in the field of cognition science. For the few analyses done for the totality of production from the three faculties 2008-2012 the duplicates were removed.

4.3 Bibliometric analysis

The results and analysis will consist of two parts. In the first part, the OA content of LUP from the years 2008 through 2012 from the three faculties is charted overall and by publication type. The calculations, tables and diagrams were made with the help of Excel.

The second part of the study concentrates on the download statistics provided by LUP for its OA content. These will be compared to the citation analysis from WoS, for each of the five years separately and as a group. This limited the analysis to items both stored OA in LUP and indexed in WoS, limiting the available data considerably, as both publications stored OA in LUP but not indexed and WoS and publications not available OA in LUP were unable to be included in the comparative analysis. Again, Excel provided the tools for creating the diagrams that allow for comparison between the download and citation data.

To assess the correlation between downloads and citations, the Pearson product-moment correlation coefficient (PPMCC) was calculated by year and faculty. The PPMCC calculates the linear correlation, or dependence, of two variables (Ejlertsson 2012, p. 227-232). The calculations result in a number between -1 and 1, with -1 being a complete negative correlation, 0 being no correlation and 1 being complete positive correlation. In the case of the correlation between downloads and citations, a number approaching -1 indicates that downloads are negatively associated with citations; as downloads go up citations go down and vice versa. A number approaching 1 indicates a positive association; downloads and citations rise and fall together. A result of 0 indicates that there is no correlation. Note that the PPMCC is a calculation of the correlation's adherence to a *linear* dependence, and does not describe the slope of that relationship or any non-linear aspects of the correlation.

4.4 Methodological limitations

The study will make use of source materials obtained freely and openly from LUP and available via subscription from WoS. This has the benefit of the results being readily reproducible by interested parties as well as comparable to previous studies. Because no sensitive material is being used in the study, ethical aspects such as protecting the confidentiality of informants in interviews

or surveys are not relevant for the study. The present study is interested in the data from a faculty/document perspective and information that could identify particular items or researchers will be left out of the analysis.

Bibliometric methods have important limitations that must be acknowledged. Any analysis relies heavily on choices made in compilation of the dataset. For example, where a paper has more than one author, a decision must be made whether to count only the first author of the paper, count all the authors equally, or assign each of the authors a "weight" based on where in the author list their name appears. Other limitations derive from the potential for bias or misinterpretation in the calculations and indices used in bibliometric analysis and these indicators must be used and applied with careful consideration (Andrés 2009, p. 121-122). This is especially true when bibliometric indicators are used in an evaluative capacity. While the current study itself does not have an evaluative aim, care has been taken to maintain transparency in the data compilation process and a critical and multi-faceted view of the scholarly communication process when interpreting the results of the study.

5. Results and Analysis

The results of the study will be presented in two parts. Under section 5.1 the OA content of the IR LUP will be examined over time and by document type. In section 5.2 downloads and citation statistics for the three faculties will be analyzed and the correlations presented.

5.1 Open access content in the institutional repository Lund University Publications

OA content was substantially represented in the IR LUP, totaling 15.4% of content in the IR from years 2008-2012. There was furthermore a substantial rise in OA available publications over the years examined, with 7.4% of research output from 2008 OA compared with 24% from 2012. To further investigate where this increase in OA content is coming from, the OA data were examined by faculty and document type. Only those document types comprising over 2% of the total content over the examined years for the three faculties are included in Figures 1-3.

Figure 1 shows the OA content indexed in LUP from FM from 2008-2012 by document type. The graph clearly shows how the percentage of OA content for these document types, with some instability, increased overall from 2008 to 2012, with the exception of the category book chapters which saw no OA content for any of the years. The total OA content for FM rises from 3.4% of content from 2008 OA to 25.7% from 2012, despite the fact that the total number of publications remained steady at around 2100 for the five years examined. For the entirety of FM's content 2008-2012 journal articles constituted 86.7%, and this is also the category that saw the most dramatic increase in percentage OA content 2008-2012, from 4.1% to 26.3%, as shown in Figure 1. For FM, OA seems to have established itself in both the publication habits of researchers and the routines of the faculty library.

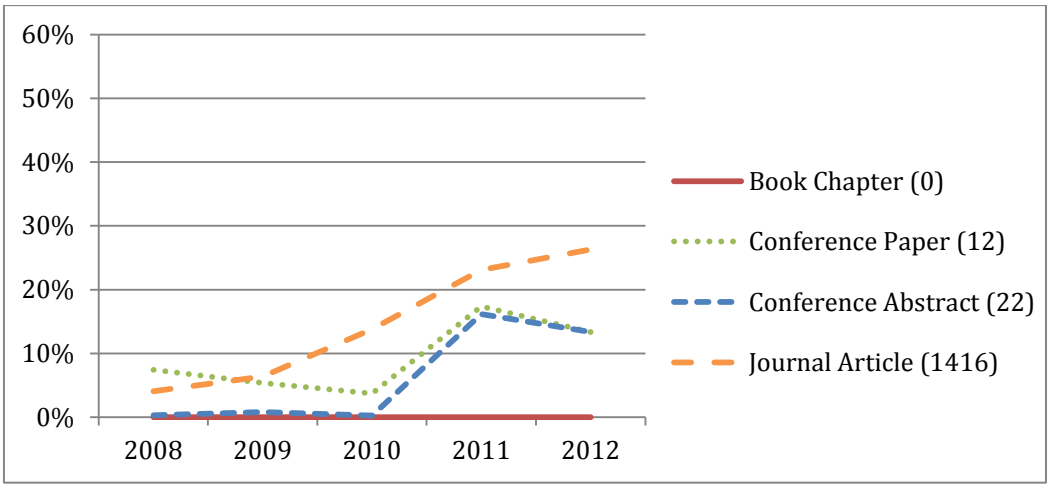

Figure 1. Percent open access content from the Faculty of Medicine 2008-2012 indexed in Lund University Publications by document type, with total amount of open access publications in parentheses

Results from FE show positive developments in the availability of OA publications from FE from the years 2008-2012. OA publications increased from 14.4% of content from 2008 to 28.1% of content from 2012. Figure 2 contains the percentage OA content for different publications types for OA content 2008-2012. Here, conference publications were the dominant OA publication type. For the totality of content from FE 2008-2012, conference papers represented 46.9%, conference abstracts 5.7% and journal articles 44.7%. However, when it comes to OA content, Figure 2 shows conference abstracts increase substantially as a source of OA content in publications from 2008-2012. In contrast, journal articles remain at disproportionally low percentage of OA content from the years 2008-2012, showing no substantial increase. FE thus shows some positive trends in percentage OA content over the five years, though there seems to be some discrepancy in representation in the totality of content and the OA content by publication type.

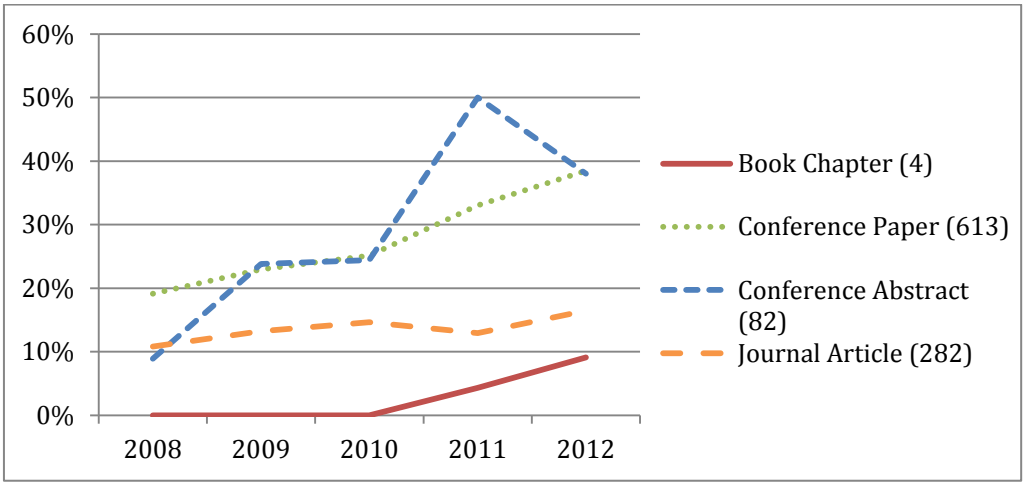

Figure 2. Percent open access content from the Faculty of Engineering 2008-2012 indexed in Lund University Publications by document type, with total amount of open access publications in parentheses

FHT's OA content in LUP showed both similar and dissimilar trends to FM and FE. Dissimilar to FM and FE, OA content in LUP for FHT did not increase in publications from the years studied, remaining between 10.5% and 13.7% of the total content. For OA and non-OA publications combined from the five years, journal articles were the dominant publication type at 44.8%, followed by conference papers at 21.2%, book chapters at 19.6% and conference abstracts at 7.0%. However, compared to the OA data from Figure 3 showing percentage OA content from FHT 2008-2012 indexed in LUP, the percentage journal articles decreases to a fairly unstable 10-15% over the five years. Similarly to FE, OA content by type was not a reflection of research output at the faculty as a whole.

A key difference between FHT and the two other faculties was the presence of book chapters as a substantial document type for both the totality of research output and for OA content. This is contrasted with FM, which was dominated by journal articles and FE, which was dominated by both journal articles and conference papers. This is not a surprising finding, as the humanities are generally thought to have a higher frequency book publications, although this is an over-simplification and there is in fact a wide variance in the use of different publishing mediums in the humanities (Hammarfelt 2012, p. 30-31). FHT seems to have a more diverse publication set compared to FM and FE, both for OA and non-OA content, though their adoption of OA seems to show instability and no overall positive trend. This may be because of either researchers not publishing OA or their OA content not being available in the IR for download because of inconsistent archiving routines. However, it is important to note that of the three faculties analyzed FHT had the least amount of total publications indexed in the IR, at 2880 items, 355 of them OA. The small data sample could be the reason for the instability in trends; a difference of a few publication items can change percentages drastically from year to year. Another factor potentially contributing to the small dataset for FHT is the exclusion of full monographs due to the "quality controlled" criteria in the compilation of the dataset. However this is unlikely to have positively affected the percentage OA content in LUP; as the report by Lindh and Wiklund (2010) showed, monograph publishers in humanities subject areas have a negative view of OA due to concerns over diminished book sales.

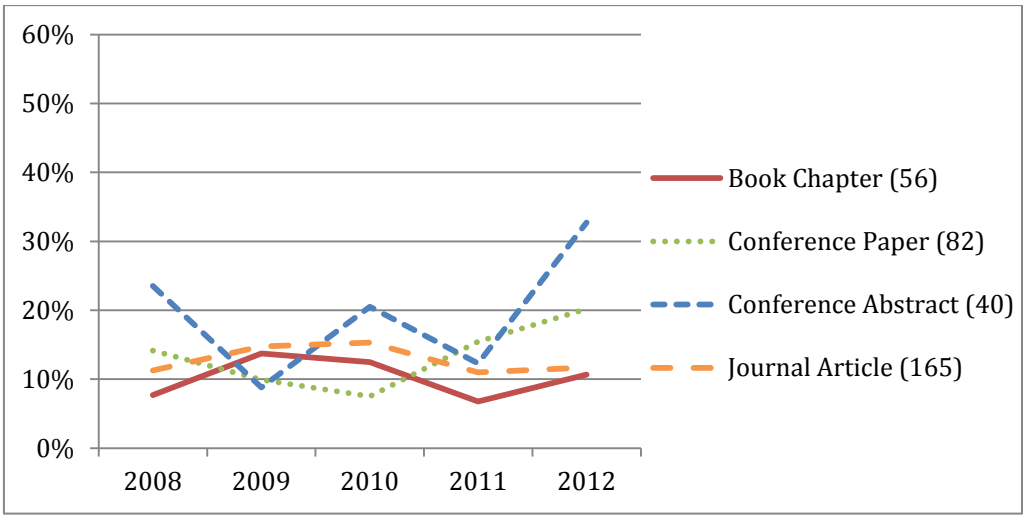

Figure 3. Percent open access content from the Faculties of the Humanities and Theology 2008-2012 indexed in Lund University Publications by document type, with total amount of open access publications in parentheses

The findings in the first part of the study show that OA content in the IR LUP is varied both for the five years examined and from the three faculties and that these divergent trends come into further focus when examined by document type. For FM, a clear positive trend in OA content was shown for publication items from the five years; their dominant publication type, journal articles, showed a dramatic increase. From this it can be inferred that both OA publishing and OA archiving routines in the IR increased for the years studied. FE also showed a positive trend in OA content; however, in their OA material journal articles were underrepresented while conference abstracts were overrepresented as a percentage of the total OA and non-OA content. FHT were more diverse overall in their publication types; however, percentage OA content did not increase in the content over the years analyzed and there was a greater degree of instability in OA publishing by type over the five years. That said, it is difficult to draw any concrete conclusions with the limited data sample from FHT.

5.2 Download statistics and citation analysis

Figure 4 shows the average number of citations from WoS and the average number of downloads from LUP for publications from 2012 to 2008 for FM, FE and FHT with years descending for the OA LUP publications indexed in WoS. The chart shows how MF leads in average number of citations during the five-year period, followed by FE and finally FHT. There is also a clear positive trend for all three faculties, with older material having been cited more than newer material, though FHT dips somewhat between 2011 and 2010. The averages from the three faculties seem to have not reached a plateau yet, with the publications continuing to accrue new citations each year over the years studied – had they done so the graph would have shown similar averages for the oldest publications. The increase remains somewhat stable during the five years for all three faculties. However, as these are averages, it may be that citations distribution changes over the years analyzed; for example, from many

publications receiving only a few citations each in the more recent years to a few articles being very well cited from the older materials – the results of the correlation analysis presented below will further illuminate these factors. However, we can see from these averages that publications from all three faculties continue to be relevant in the academic community.

The figures are less clearly interpreted for the download statistics. There is a slight positive trend under the five years, with older materials continuing to accrue downloads, but this trend is less pronounced than with citations. This seems to indicate that downloads may more quickly reach a stable level. FE leads in number of downloads, followed by FHT and FM. Looking at the two graphs in Figure 4 together, it is clear that citations and downloads do not follow the same trend. However, it is important to take into consideration that the dataset from FHT was very small: as little as 5 publications in 2008 with both download and citation statistics available. This could account for the variability in download averages – one very frequently or infrequently downloaded item could easily skew the dataset.

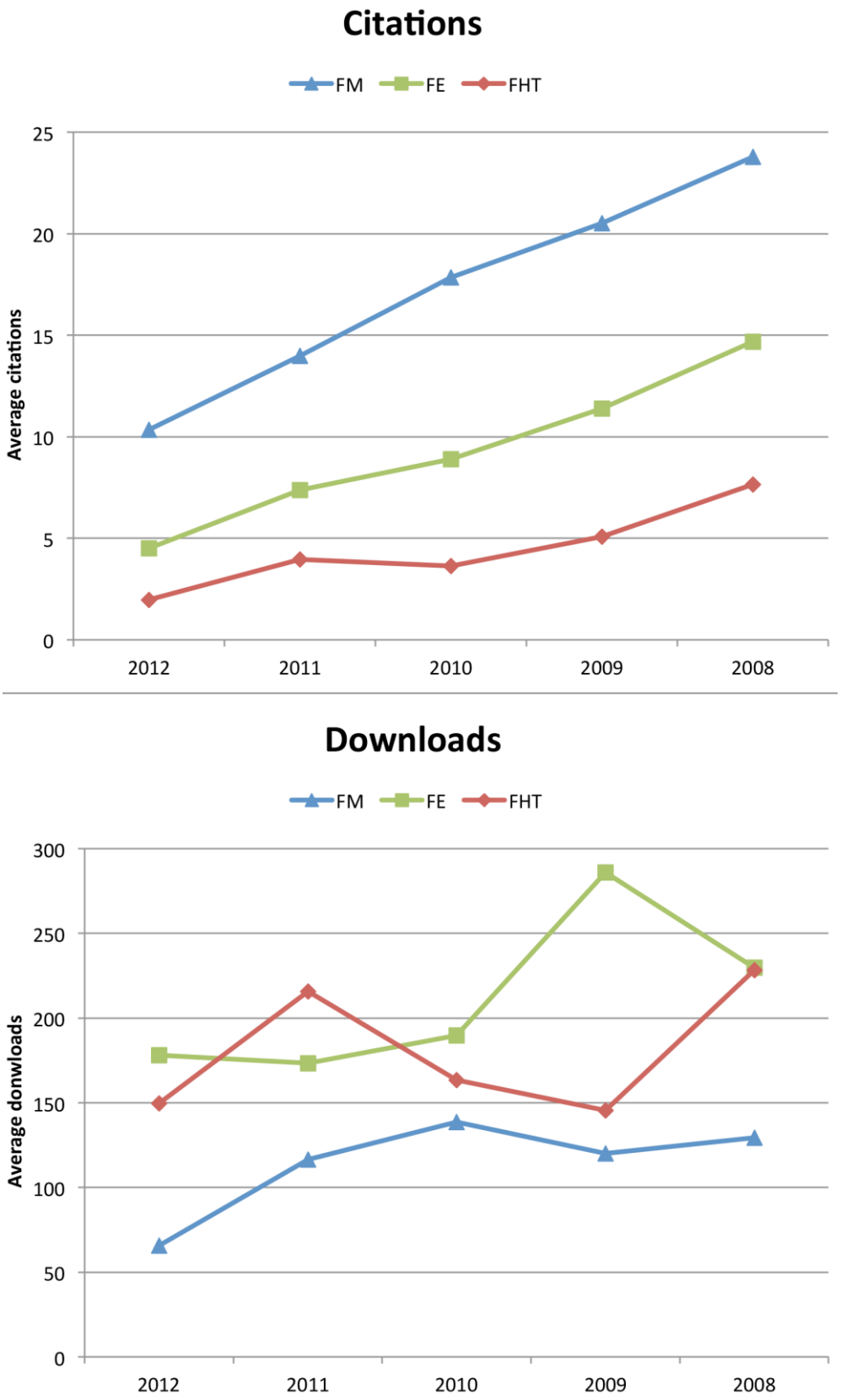

Figure 4. Average number of citations from Web of Science and average number of downloads for publications from Lund University Publishing 2008-2012 for the Faculty of Medicine, the Faculty of Engineering and the Faculties of the Humanities and Theology with years descending for open access publications indexed in Web of Science

Figure 5 shows the results of the correlation between the download statistics in LUP and the citation analysis in WoS for each of the five years. From the graphs we can clearly see a wide spread in the citation and download data. However most of this spread is accounted for by isolated outliers; the majority of the data was characterized by comparatively few downloads and citations. The PPMCC calculations are given for each faculty and year, with none of the correlations being above 0.5 or below 0.1, revealing weak correlations. The total number of publications in each dataset is also given to show how much data was available for the correlation calculations. Any finding of correlation must also be tempered with the fact that these graphs represent publication *from* each year and not *year by year*. For instance, it cannot be shown with these results whether a publication first is downloaded many times and then receives citations, receives some citations first and is then downloaded, or if any development in citations and downloads increases at about the same rate for any given document. For such an analysis both the citation and download data for the publications would need to be broken down year by year – this data was not readily available for the study. Furthermore, only the oldest publications could then be analyzed, excluding much of the more recent OA content.

For FM, a weak positive correlation was shown for the oldest publications, which could suggest a slight correlation between downloads and citations. However, looking at the scatter plot, we can see that a few publications are being downloaded or cited many times, and these publications are generally not concomitant. Especially notable is that publications in all years except 2010 receive comparatively great numbers of citations, perhaps an indication of the Matthew effect or the products of especially esteemed authors. There are also from all years except 2009 one or more publications outliers on the download axis, that is to say, from each year there are a few select publications that are being downloaded to a markedly greater extent than other publications from FM that year.

FE's correlation calculations again show weak correlations, which, conversely to what one might expect, become weaker for the older publications. In other words, downloads become less and less correlated with citations as the publication ages, eventually showing a very slight negative correlation for the publications from 2008. Downloads and citations represent more and more separate variables as time goes on. Analyzing the scatter plots, this may be because of a few older publications being heavily cited, with one publications from 2008 having over 150 citations compared to the next most cited having only just over 50 citations.

The correlation analysis for FHT shows no substantial correlation for 2009-2012, and a weak correlation for 2008. The number of items for FHT was however so small as to make any conclusions difficult to draw. In addition, while many of the publications were being substantially downloaded, citation activity for these publications was overall low. One reason for the small dataset could be the absence of monographs from the dataset, due to monographs not being included in the "quality controlled" category in LUP which was a criteria

in the compilation of the dataset and low coverage of monographs in WoS. Nevertheless, the main finding for the FHT correlation analysis is that this type of analysis is ill suited to application in the humanities subject fields, as the number of publications indexed in both LUP and WoS is simply too small to draw any conclusions from.

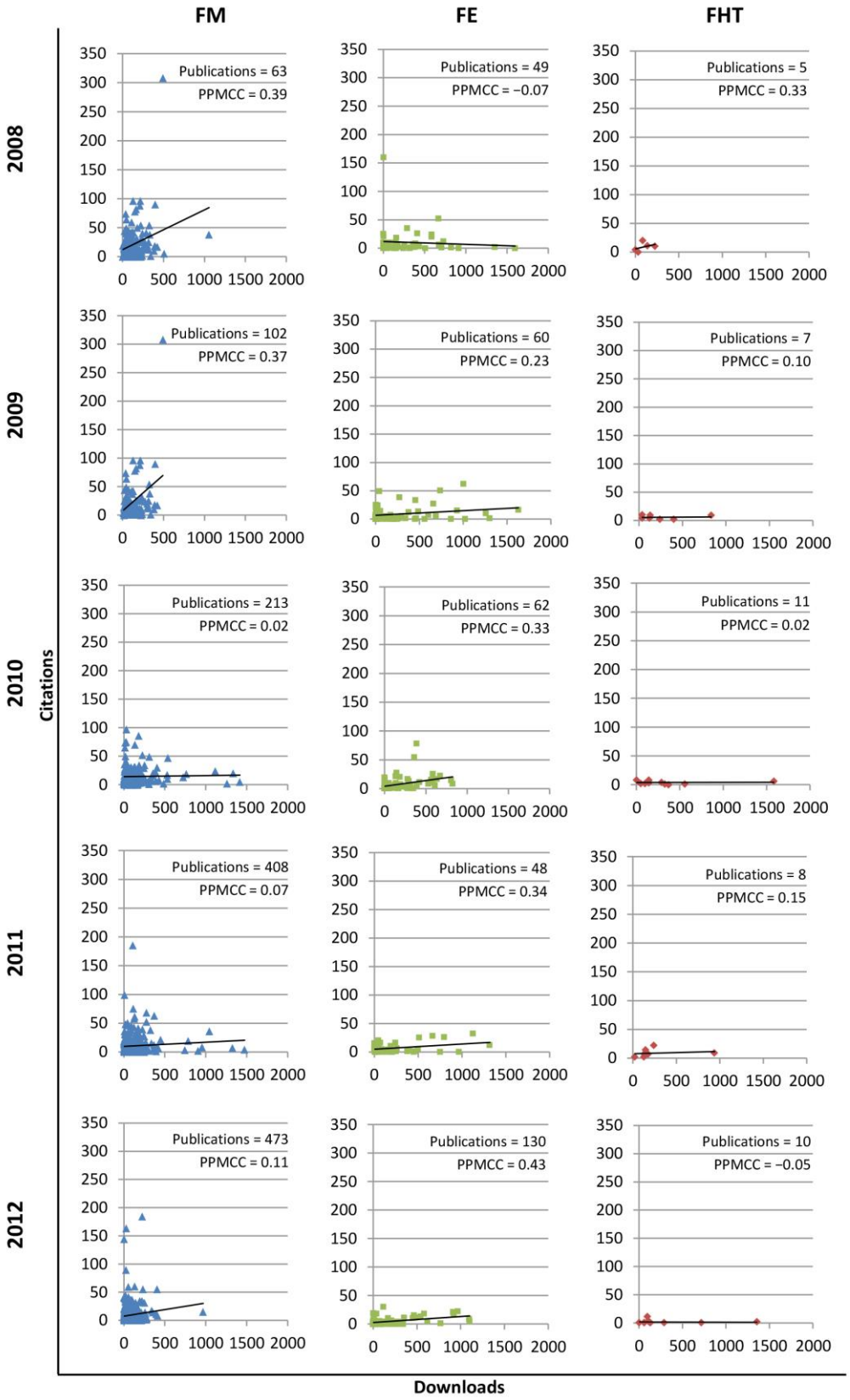

Figure 5. Citations in Web of Science (WoS) as a function of downloads from Lund University Publishing (LUP) 2008-2012 from the Faculty of Medicine, the Faculty of Engineering and the Faculties of the Humanities and Theology, with total number of publications in LUP with WoS numbers per faculty and year and Pearson product-moment correlation coefficients (PPMCC)

6. Discussion

The results of the study show a complex picture of the way that OA manifests itself in the IR landscape. For the discussion part of the thesis, first the results and analysis for each of the faculties will be discussed, with emphasis on the publishing characteristics of OA compared with non-OA. The results from each faculty will be tied together and analyzed within the framework of Whitley's (2000; 2008), Fry and Talja's (2007) and Merton's (1968; 1973; 1988) theories and in relation to previous research. The second part of the discussion will further discuss the implications of the results of the analysis of the download statistics and download statistics' potential in scientific evaluation.

6.1 Scientific organization of open access content in Lund University Publications

OA publishing in LUP increased substantially for publications indexed from 2008 to 2012, from 7.4% in 2008 to 24.0% in 2012. It can therefore be concluded that OA publishing in the IR has successfully been integrated into the university's research output routines. However, analyzing OA content by faculty and document type revealed that trends in OA publishing varied between the faculties. This reflects, as the following discussion elucidates, differences in the work organizations of the different subject fields the faculties are comprised of.

FM exhibited clear tendencies in its publishing patterns and trends. OA publications increased dramatically from a few percentage points in publications from 2008 to almost a quarter of publications 2012. This is likely due to a variety of factors working together, such as improved library routines for archiving OA material and, probably most importantly, increased pressure from research funders, such as the Swedish Research Council, to publish OA. OA publishing and archiving has demonstrated itself conclusively in the IR. Journal articles were the predominant medium of scholarly communication for the research output from the faculty, and this is where the percentage of OA content increased the most for publications from 2008-2012. Borrowing from Whitley's (2000) concept of audience in scholarly communication, we can infer from FM's preference for scientific, peer-reviewed journals that they have a low diversity of audience for their scholarly communication. According to Whitley's theory this is a sign of consensus on research goals and procedures. FM was the most productive faculty from 2008 to 2012, having had the most publications. Citations to the faculty's publications were also high – researchers appear to be dependent on each other's findings for the meaningful

interpretation and presentation of their results. The findings seem to indicate an overall high mutual dependence and low task uncertainty in FM, though this of course likely varies substantially between the disciplines and subject fields represented at the faculty.

The publishing patterns at FE showed a greater degree of diversity than FM, indicating that FE potentially has greater audience diversity than FM, and consequently a potentially greater plurality of research goals and procedures. Conversely, citation activity was high at the faculty, indicating that researchers are still dependent on the work of their colleagues in producing results. These combined factors indicate a field with both high mutual dependence, as authors cite and use one another's work, and high task uncertainty, given the plurality of audiences. However, these results may be unreliable given the potential for inconsistent application of the document type category "conference abstract", an umbrella term which encompasses a wide range of publications. For example, the texts can be anything from short abstracts describing a lecture to conference posters.

Similarly to FM, FE had a sharp increase in the amount of OA publications. However for this faculty the predominant publishing medium, journal articles, was not the source of the main increase in OA material. The greatest portion of this increase came instead from conference abstracts. Here, OA, instead of coming into the already established publishing medium, is manifesting itself in another communication form. This could be due to a variety of factors, for example pressures from funding bodies, changes within the scholarly field, or changes in library IR indexing routines. Without further qualitative investigation, definitive conclusions cannot be drawn. Again, it is hard to interpret these results, as the category of "conference abstracts" is a broad one with many differing document types under this umbrella term.

Of the three faculties FHT had the most diversity in publishing types. According to Whitley's theory on audience structure's influence on scientific fields, this would encourage diversity of research goals and methodology. This diversity is further enhanced by low citation behavior, indicating a field whose dependence on the work of their colleagues to make authoritative knowledge claims is low. From the available data and analysis it would seem then that the subject fields in FHT are characterized by a tendency towards low mutual dependence and high task uncertainty. These results are coupled with the results of the analysis of OA content, which showed that OA content did not increase over the years analyzed. A major contributing factor to this result could be that the Swedish Research Council's OA policy excludes books and book chapters, with book chapters, as the results have shown, being an important publishing medium at FHT. It is also important to note here that due to the "quality controlled" criteria in the data compilation process, monographs were excluded from the dataset, and as discussed in the results and analysis section of the thesis, in light of Lindh and Wiklund's (2010) report noting that monograph publishers in the humanities were reluctant to embrace OA, the addition of books to the dataset was unlikely to increase the percentage OA for FHT. In addition, journal articles were also an important publishing medium

and they did not show an increase in percentage OA content in publications from the five years either.

We have seen from the results, analysis and this discussion that multiple factors are contributing to the successful integration of OA publishing in LUP, which leads to varied trends in OA content in the faculties examined in this thesis. For an overreaching explanation for the differing trends, Fry and Talja (2007) provide a possible explanation. They noted that in implementing new communication models, the new medium must adhere to existing standards in the subject field in order to be successfully integrated (ibid, p. 131). This may explain the success of OA in FM compared to FHT, as FM has a more homogeneous publication type, simplifying OA implementation in publishing procedures. This cannot, however, explain the success of OA in FE, as the increase was in a new publication type, conference abstracts, compared to their overall main publication medium being journal articles. It does, however, bring up an interesting point in that OA must work with existing research practices to be successful.

6.2 Download statistics and the evaluation of the sciences

The results of the citation analysis are in line with generally accepted patterns revealed by bibliometric studies of scientific communication: the medical-oriented subject fields received the most citations and the humanities-oriented fields the least, and citations steadily grew as publications aged (with the exception of 2011-2010 for FHT, which saw a very slight decrease in citations). The download statistics, however, showed a differing pattern, with FE and FHT leading, followed by FM. These trends were however less stable than with the citation analysis, and the fact that the data set for FHT was very small could be further contributing to the instability of the download trends from this faculty. Any direct correlation between download statistics and citation analysis remains uncertain after the correlation analysis performed in the study. No clear correlation presented itself for FM and FE and the amount of publications from FHT was too few to draw any conclusions from. Most publications were cited and downloaded few times, with some isolated publications being either cited or published many times. Therefore, the results indicate that citations and downloads represent distinct phenomena.

Taken from Merton's (1973) theory of citations as a type of science capital reward, the total citation counts of FM, FE and FHT represent these faculties' total generated science capital over the course of the examined five years and thus the results of the citation analysis alone would be a good measure of research impact for the faculty. This, however, is a grave simplification, especially with regards to the citations analysis for content from FHT. There could be many explanations for FHT's low average of citations that are not tied to the impact of the faculty's research output. One interpretation of the citation analysis results is that a low level of mutual dependence between researchers leads to a diminished need to relate one's work to others' in the field, reducing citation behavior, as related to Åström and Sándor's (2009) citation model,

where the humanities fields lean more towards a negotiating or distinctive citation model rather than a cumulative one, decreasing the need to for heavily citing colleagues. However, further factors are likely contributing to the lower average citations for FHT. One important factor is that the humanities publish a greater portion of their work in languages other than English (Hammarfelt 2012), which decreases their visibility in WoS and lowers the portion of research output available for citation analysis. The humanities generally also publish less frequently, meaning that they have less opportunity to cite and be cited. Another factor could be that for many of the subject fields in the humanities there is not a strong "research front", which would lead to a preference to cite older sources.

The results of the citation analysis show furthermore that many scholarly publications will receive few or no citations compared with a few that receive considerably more attention. This is a finding common to bibliometric analysis and is in line with the inflation of attention already described by Merton's Matthew effect (Merton 1968). That this could be manifesting itself in the download statistics as well is a more interesting finding, as one explanation for increases in citations is that the citation serves to bring a scholarly publication to the attention of other researchers, increasing an item's prestige. The download and use of an article does not share this mechanism, being a usage measure that is less readily visible by itself, which means that the inflation could be due to downloads being correlated with other means of sharing scholarly consumption, for example informal channels of researchers mailing each other article links or sharing a publication on social media. Thus, downloads would indirectly increase an article's exposure within the scholarly community.

The results of the citation analysis together with the download statistics build a more complex picture of research impact at the faculties. FE and FHT lead with regards to average downloads per document, while FM trails behind, despite the fact that FM had shown a substantial increase in the percentage of content available OA. This may be a function of FE and FHT having a wider audience than FE, which would be in line with the finding that FE and FHT both have a more diverse publishing base than FM, indicating a wider audience for their research output. The download statistics have the potential to capture the use and impact of scholarly resources by audiences outside academia. In addition, this kind of usage data is inevitably tied to information retrieval. In the cases where OA is a part of parallel publishing and a publication item is available in both a licensed database accessible for example through the university and through an OA IR, the "use" of the article will be counted either in the publisher's database or through the IR. These channels may represent different user groups, for example academics preferring to use the publisher's database or their university's discovery search tool, whereas a member of the general public might find the same item by searching Google and arriving at the IR's website.

While finding that download statistics presented a way to measure the impact of scholarly output for users outside academia would be exciting, more research

is needed to investigate the meaning of download statistics before these can be used as a measure of research impact in any evaluative capacity (see discussion under section 2.3.2). What potentially makes this avenue of inquiry particularly relevant is increased interest in measuring research impact in a broader context. The Swedish Research Council in their latest report have a separate category for research impact outside academia in their model for resource allocation though performance evaluation (Swedish Research Council 2014, p. 11), indicating a desire to capture and evaluate the impact of research outside academia by research funders.

Another potential use of download statistics from an IR would be their use as an early predictor of citation, being a measurement of impact that takes less time to establish itself than citations. Citations take considerable time to accumulate post-publishing given that the citing works themselves will need to be published. However, if downloads were an early predictor of citations, we would expect to see citations more strongly correlate with downloads the farther back in time one goes. High download counts from 2012 would slowly give way to a spread toward the middle of the graph as citations caught up with downloads. The only faculty where this might be the case was FM, as correlation increased weakly as time wore on. However, the opposite was true of FE, where correlations weakened as publications aged. Therefore, it might be worth further investigation for the medical fields if downloads could be used as an early predictor of citations; the very weak correlations from this study are simply not robust enough to draw any conclusions from, but a similar study to the present one in a few years' time might reveal correlations from FM continuing to strengthen.

The results of the study seem to be in consensus with Kurtz and Bollen (2011): the relationship, if any, between usage of scholarly output and citations is a complicated one, whose meaning is at the moment unclear. As Kurtz and Bollen point out, "The fact that the universe of users of scholarly articles can be much broader and different from the universe of scholarly authors presents both substantial challenges and substantial opportunities" (ibid, p. 19). The results of this study have shown that download statistics have the potential to provide a unique measurement of the usage of scholarly output by being a distinct measurement from citation analysis. Their meaning and relationship to citation analysis may however vary between subject fields, and more research is needed before usage data can be usefully applied to research impact evaluation.

7. Conclusions

This thesis has aimed at investigating the role of IRs in OA publishing and examining the potential IRs in describing research impact through the download statistics they provide for their OA material. The study has asked the following questions:

- What does the OA landscape look like in an IR and what can IR metadata, download statistics and citation analysis together reveal about the organization of work in scholarly research output in relation to OA publishing?

- How do the download statistics of OA publications in an IR compare with WoS citation analysis and how can download statistics from an IR inform our understanding of research impact?

The present thesis has thus examined the OA landscape of IRs from two perspectives, the first being a descriptive examination of the data from the IR and the second being an investigation of the possibilities provided by download statistics in describing the impact of OA content in IRs.

The descriptive analysis showed that the percentage of OA content in the IR LUP increased during the years 2008 to 2012, with a greater portion of the newer publications being OA. However, this was not true for all faculties examined in the study; OA publications increased sharply for FM and FE while FHT's percentage OA publication remained stable. The downloads statistics and citation analysis of the OA content in LUP showed that while citation data climbed steadily with older publications having more citations than recent ones, the download statistics were less stable in their trends. Furthermore, there was no substantial correlation between downloads and citations for the examined years, showing that downloads are measuring impact from a different user group than citations are. However, the dataset from FHT was very small and instability in the download/citation numbers may be due to insufficient sample size.

Using the theory put forth by Whitley (2000) and expanded on in relation to digital scholarly resources by Fry and Talja (2007) the results and analysis have shown that the organization of scholarly communication produced by these faculties is dissimilar, which points towards differing levels of mutual dependence and task uncertainty. With the domination of one publication channel, journal articles, and the high level of citations, FM showed high mutual dependence and low task uncertainty. FHT, having greater spread of

publishing channels and exhibiting relatively lower citations, revealed itself to be a field with lower task uncertainty and higher mutual dependence. These factors also indicate that publications from FHT are aimed at a wider audience than FM. These differences in work organization may explain why FM has had greater success in implementing OA.

With regards to the download statistics' potential for evaluation of research impact, the results show that download statistics do not correlate with citation analysis and that download statistics are therefore a novel measurement of research impact. More research is need to investigate the potential meaning of these download statistics, particularly in relation to their potential to reveal scholarly output's impact on audiences outside academia. In addition, and in line with previous research by Hammarfelt (2012), research evaluation in the humanities should make use of methodology and measurements that take into consideration the characteristics of humanities' publishing behavior, given that publishing behavior at FHT differed substantially from the other faculties examined, and that this faculty overall had very little data available for the comparative analysis given that its research output was poorly represented in WoS.

In summation, the descriptive analysis of data from the IR LUP pointed to a heterogeneous OA landscape in the IR, with variations between the faculties in OA trends and content types while the download and citation correlation revealed that download statistics represent a unique indicator of impact from citation analysis, enabling the measurement of a potentially broader user group. Thus, the thesis's research questions have been answered. Perhaps most importantly, IRs' role in making OA publications available has been illustrated and IRs' potential as a source of valuable data on the organization of the sciences has been confirmed.

OA publishing remains as of the writing of this thesis in a state of instability. Researchers, publishers, financiers, libraries and others all have a stake and are negotiating the future of scholarly communication. Elsevier, a major publisher of scholarly journals, has as of April 2015 updated it policies to clarify its position on OA and scholarly sharing (Wise 2015). It particularly brings up IRs right to make Elsevier-journal published material freely available. Reactions critical to these policy changes have already found their way to blogs, claiming that Elsevier's policy changes represent a step backwards from increasing authors' freedom to openly share their work (Bremsbs 2015; Harnad 2015). IRs stand in the middle of a tug-of-war in scholarly communication; the key for success in an uncertain climate is for academic libraries to keep apprised of changes, maintain optimal flexibility and above all find ways to best exploit the tools they have – including IRs – to the advantage of the researchers at their institutions.

7.1 Further research

The dataset analyzed for the present study has the possibility of providing many more results than were presented here. Only those results that contributed to

answering the thesis's research questions were presented. Any further analysis will be left to future studies. For example, the dataset collected for the study could be used to investigate whether OA content was more cited than non-OA content, expanding on Norris, Oppenheim and Rowland's (2008) work.

The results of the study presented in this thesis highlight the importance of ongoing research into the IRs, OA, the work organization of the sciences and the way changes in these might affect the evaluation of the sciences. By finding varying results for the different academic areas studied, the study in this thesis has reiterated the need for careful consideration of the subject field when using quantitative methods in evaluations of science. In addition, further studies will be needed to investigate academic libraries' approaches to OA in IRs and how these might vary across different academic fields. One area in particular the study had uncovered a need for further research into is an examination of the consumption of OA scholarly literature in IRs in order to understand whether download statistics can be used to investigate the impact of research on audiences outside academia. For instance, research is needed to investigate how users arrive at IR websites and if this affects their behavior in downloading content from the IR.

IRs have and will continue to play a part in the development of alternatives in OA publishing, and developments are awaited in the next few years, as with the case of SwePub. These developments will necessitate periodical reviews that take both a practical and theoretical approach to IRs. In particular, the study has highlighted the need for investigation into the role of academic library policy in the shaping of IRs and how practicing librarians tackle OA publishing developments. The present thesis has reiterated the potential for IRs to provide quality data for analysis of academic subject fields and the impact of research. However, this must be complimented by qualitative studies that examine researchers' motivations in searching and consuming scholarly information, as well as in citing and publishing their own work, in order to paint a complete picture of IRs' current role and in order to fulfill IRs' potential in scholarly communication. It is critical to not become too dazzled by the statistics that quantitative studies can provide and remember that bibliometrics is but one tool in the library and information scientist's toolbox for examining scholarly communication phenomena.

8. References

Andrés, A. (2009). *Measuring academic research: how to undertake a bibliometric study*. Oxford: Chandos.

Bailey, C. (2007). What Is Open Access? *Digital Scholarship*. http://digital-scholarship.org/cwb/WhatIsOA.htm [2015-03-21].

Bailey, C. (2008). Open access and libraries. *Collection Management*, 32(3-4), p. 351–383.

Bakkalbasi, N., Bauer, K., Glover, J. & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 3.

Baldi, S. (1998). Normative versus Social Constructivist Processes in the Allocation of Citations: A Network-Analytic Model. *American Sociological Review*, (6), p. 829.

Bar-Ilan, J. (2008). Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), p. 257.

Bar-Ilan, J. (2010). Citations to the 'Introduction to informetrics' indexed by WOS, Scopus and Google Scholar. *Scientometrics*, 82(3), p. 495–506.

Bharathi, D. (2013). Methods employed in the Web of Science and Scopus databases to effect changes in the ranking of the journals. *Current Science (00113891)*, 105(3), p. 300–308.

Björk, B. & Solomon, D. (2014). How research funders can finance APCs in full OA and hybrid journals. *Learned Publishing*, 27(2), p. 93–103.

Bonilla-Calero, A. (2014). Institutional repositories as complementary tools to evaluate the quantity and quality of research outputs. *Library Review*, 63(1/2), p. 46–59.

Bonilla-Calero, A. (2008). Scientometric analysis of a sample of physics-related research output held in the institutional repository Strathprints (2000-2005). *Library Review*, 57(9), p. 700–721.

Borgman, C. (2007). Scholarship in the digital age: information, infrastructure, and the Internet. MIT Press, Cambridge, MA.

Bornmann, L. (2014). Measuring the broader impact of research: The potential of altmetrics. http://arxiv.org/ftp/arxiv/papers/1406/1406.7091.pdf [2015-03-22]

Bremsbs, B. (2015). Is this supposed to be the best Elsevier can muster? bjoern.brembs.blog. http://bjoern.brembs.net/2015/05/is-this-supposed-to-be-the-best-elsevier-can-muster/ [2015-05-19]

Bradford, S. (1985). Sources of information on specific subjects 1934. *Journal of Information Science,* 10(4), p. 176–180.

Cho, J. (2014). Intellectual structure of the institutional repository field: A co-word analysis. *Journal of Information Science,* 40(3), p. 386–397.

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the Science citation index to cybermetrics*. Lanham: Scarecrow Press.

De Bellis, N. (2014). History and Evolution of Biblio(Metrics). In: Cronin, B. & Sugimoto, C. (eds.), *Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact*. Cambridge: MIT Press, p. 23-44.

Delgado, E. & Repiso, R. (2013). The Impact of Scientific Journals of Communication: Comparing Google Scholar Metrics, Web of Science and Scopus. *El impacto de las revistas de comunicación: comparando Google Scholar Metrics, Web of Science y Scopus.*, 21(41), p. 45–52.

Dorner, D. and Revell, J. (2012). Subject librarians' perceptions of institutional repositories as an information resource. *Online Information Review*, 36(2), p. 261–277.

Dutta, G. & Paul, D. (2014). Awareness on Institutional Repositories-related Issues by Faculty of University of Calcutta. *DESIDOC Journal of Library & Information Technology,* 34(4), p. 293–297.

Ejlertsson, G. (2012). *Statistik för hälsovetenskaperna*. Studentlitteratur: Lund.

Engwall, L. & Nybom, T. (2008). The Visible Hand Versus the Invisible Hand: The Allocation od Research Resources in Swedish Universities. In: J. Gläser and R. Whitley, eds., *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems,* Sociology of the Sciences Yearbook: 26. Dordrecht: Springer Science+Business Media B.V., p. 31–49.

Faculty of Medicine (2014a). *Parallellpublicering*. http://www.med.lu.se/intramed/forska_utbilda/publicera/parallellpublicering [2015-04-01]

Faculty of Medicine (2014b). *Vad är LUP?*. <http://www.med.lu.se/intramed/forska_utbilda/publicera/lup [2015-04-01]

Fry, J. & Talja, S. (2007). The intellectual and social organization of academic fields and the shaping of digital resources. *Journal of Information Science*, 33(2), p. 115–133.

Furner, J. (2014). The Ethics of Evaluative Bibliometrics. In: Cronin, B. & Sugimoto, C. (eds.), *Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact*. Cambridge: MIT Press, p. 85-108.

Garfield, E. (1955). Citation Indexes for Science. *Science*, (3159), p. 108.

Garfield, E. (1979). *Citation indexing: its theory and application in science, technology, and humanities*. Information sciences series. New York: Wiley.

Gilbert, G. (1977). Referencing as Persuasion. *Social Studies of Science*, 7(1), p. 113–122.

Gosnell, C. (1944). Obsolescence of Books in College Libraries. *College & Research Libraries*, 5(2), p. 115–125.

Graffner, M. (2014). *LUP info*. https://lup.lub.lu.se/lupInfo?func=loadTemplate&template=0about [2015-01-23]

Gross, P. & Gross, E. (1927). College Libraries and Chemical Education. *Science*, 66(1713).

Hammarfelt, B. (2012). *Following the Footnotes: A Bibliometric Analysis of Citation Patterns in Literary Studies*.

Hammarfelt, B., Nelhans, G., Eklund, P. & Åström, F. (2015). The heterogeneous landscape of bibliometric indicators: Evaluating models for allocating resources at Swedish universities. *Reseach Evaluation*. [pre-print]

Hammarfelt, B. & de Rijcke, S. (2014). Accountability in context: effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the faculty of Arts at Uppsala University. *Research Evaluation*, 24(1), p.63.

Harnad, S. (2015). Elsevier updates its article-sharing policies, perspectives and services. Open Access Archivangelism. http://openaccess.eprints.org/index.php?/archives/1150-Elsevier-updates-its-article-sharing-policies,-perspectives-and-services.html [2015-05-19]

Haustein, S. (2014). Readership Metrics. In: Cronin, B. & Sugimoto, C., (eds.) *Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact*. Cambridge: MIT Press, p. 327-344.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H. & Terliesner, J. (2013). Coverage and adoption of altmetrics sources in the bibliometric community. http://arxiv.org/ftp/arxiv/papers/1304/1304.7300.pdf [2015-03-24]

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2), p. 251–261.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature*, 520(7548), p. 429–431.

Inefuku, H. (2013). Whatever Happened to Art and Design?: Using Archival Practice to Manage the Impact of Academic Restructuring on Institutional Repositories. *Journal of Library Administration*, 53(4), p. 209–222.

Jones, C. (2007). *Institutional repositories: content and culture in an open access environment*. Chandos information professional series. Oxford: Chandos.

Kurtz, M. & Bollen, J. (2011). Usage Bibliometrics. http://arxiv.org/abs/1102.2891 [2015-03-24]

Latour, B. & Woolgar, S. (1979). *Laboratory life: the social construction of scientific facts*. Sage library of social research, 80. Beverly Hills: Sage.

Lindh, K. & Wiklund, G. (2010). *Open access för humanister och rättsvetare. En kartläggning av publiceringspolicy och praxis inom nordisk utgivning. Slutrapport*. BIVIL:s skriftserie. Lund University, Department of Cultural Sciences. https://lup.lub.lu.se/search/publication/1578796 [2015-04-08]

Lotka, A. (1926). Statistics-The Frequency Distribution of Scientific Productivity. *Journal of the Washington Academy of Sciences*, 16(12). http://listserv.utk.edu/cgi-bin/wa?A3=ind0709&L=sigmetrics&P=52661&E=2&B=--%3D-YUefx%2F0auEG8%2B29U7Cdc&N=Lotka+1929.pdf&T=application%2Fpdf [2015-003-19]

Lund University Libraries (n.d.). *LUP Search User Guide*. https://lup.lub.lu.se/search/doc/userguide#sec-3-6 [2015-04-07]

LUP Statistics. (n.d.). http://lup.lub.lu.se/lupStat?tab=5_total_top_publications&org=_all&type=nosp [2015-03-11]

McCain, K.W. (2014). Obliteration by incorporation. In: Cronin B. and Sugimoto, C. (eds.), *Beyond bibliometrics : harnessing multidimensional indicators of scholarly impact*. Cambridge: MIT Press, p. 129-150.

Meho, L. & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science & Technology*, 58(13), p. 2105–2125.

Merton, R. (1968). The Matthew Effect in Science. *Science*, (3810), p. 56.

Merton, R. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.

Merton, R. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis*, (4), p. 606.

National Library of Sweden (2014). *Lägesrapport Swepub Analys*: https://drive.google.com/file/d/0B9caWakO_ZA0THhrQzVUQUxtdW8/view?usp=embed_facebook [2015-04-14]

Neuhaus, C. & Daniel, H. (2008). Data sources for performing citation analysis: An overview. *Journal of Documentation*, 64(2), p. 193–210.

Norris, M., Oppenheim, C. & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science & Technology*, 59(12), p. 1963–1972.

Oppenheim, C. (2008). Electronic scholarly publishing and open access. *Journal of Information Science*, 34(4), p. 577–590.

Price, D. (1965). *Little science, big science*. New York: Columbia University Press.

Priem, J., (2014). Altmetrics. In: Cronin B. & Sugimoto, C. (eds.), *Beyond bibliometrics : harnessing multidimensional indicators of scholarly impact*. Cambridge: MIT Press, p. 263-288.

Priem, J., Groth, P. & Taraborelli, D. (2012). The Altmetrics Collection. *PLoS ONE*, 7(11). http://ludwig.lub.lu.se/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edselc&AN=edselc.2-52.0-84868326485&site=eds-live&scope=site [2015-01-23].

Priem, J., Taraborelli, D., Groth, P. & Neylon, C., (2010). altmetrics: a manifesto. altmetrics.org. http://altmetrics.org/manifesto/ [2015-03-22]

Small, H. (1978). Cited Documents as Concept Symbols. *Social Studies of Science*, (3), p. 327.

Stanton, K. & Liew, C. (2012). Open access theses in institutional repositories: An exploratory study of the perceptions of doctoral students. *Information Research*, 17(1).

Swedish Research Council (2014). *Forskningskvalitetsutvärdering i Sverige – FOKUS*. Stockholm: Vetenskapsrådet.

Swedish Research Council (2015). *Open access – Vetenskapsrådet*. http://www.vr.se/inenglish/researchfunding/applyforgrants/generalconditionsforgrantapplications/openaccess.106.5adac704126af4b4be280007766.html [2015-04-14]

White, H. (2004). Reward, persuasion, and the Sokal hoax: a study in citation identities. *Scientometrics*, 60(1), p. 93–120.

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford: Oxford University Press.

Whitley, R. (2008). Changing Governance of the Public Sciences: The Consequences of Establishing Research Evaluation Systems for Knowledge Production in Different Countries and Scientific Fields. In: Whitley, R. & Gläser, J. (eds.), *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*. Dordrecht: Springer Science+Business Media B.V.

Wise, A. (2015). Unleasing the power of academic sharing. Elsevier Connect. http://www.elsevier.com/connect/elsevier-updates-its-policies-perspectives-and-services-on-article-sharing [2015-05-19]

Zipf, G. (1965). *The psycho-biology of language: an introduction to dynamic philology*. MIT paperback series, 99-0115565-7; 38. Cambridge: MIT Press.

Åström, F. & Sándor, Á. (2009). *Models of Scholarly Communication and Citation Analysis*.