The deliverance of open access books : examining usage and dissemination

Snijder, R.

6 Patterns of information : Clustering books and readers in open access libraries

6.1 Introduction

Open access libraries operate in a continuum between two distinct organisation models: online retailers versus 'traditional' libraries. Online retailers such as Amazon.com are successful in recommending additional items that match the specific needs of their customers. The success rate of the recommendation depends on knowledge of the individual customer: more knowledge about persons leads to better suggestions. Thus, to optimally profit from the retailers' offerings, the client must be prepared to share personal information, leading to the question of privacy.

In contrast, protection of privacy is a core value for libraries. The question is how open access libraries can offer comparable services while retaining the readers' privacy. A possible solution can be found in analysing the preferences of groups of like-minded people: communities. According to Lynch (2002), digital libraries are bad at identifying or predicting the communities that will use their collections. It is however our intention to explore the possibility to uncover sets of documents with a meaningful connection for groups of readers – the communities. The solution depends on examining patterns of usage, instead of storing information about individual readers.

This paper will investigate the possibility to uncover the preferences of user groups within an open access digital library using social networking analysis techniques.

6.2 Background

Recommender systems are powerful tools, whose design poses privacy issues. The role of privacy in the library landscape is discussed, along with the use of recommendation systems in libraries. If it is not feasible to match titles to individuals, the use of clustering techniques might mitigate some of the privacy problems while still creating relevant sets of titles. In turn, these sets may be used in recommendation services.

6.2.1 Recommender systems

Recommender systems can be defined as tools that provide suggestions about items that may prove valuable to a user (Linden, Smith, & York, 2003; Pazzani & Billsus, 2007; Ricci, Rokach, Shapira, & Kantor, 2011; Schafer, Konstan, & Riedl, 1999). The prediction is based on processing data about items, users and transactions. Items are the objects to be recommended; in the case of digital libraries this would be documents. Understanding users is a critical part of recommender systems; ultimately, their success is based on how well they know the user's needs and preferences. Needless to say, this poses privacy issues. Transactions are defined as a recorded interaction between a user and the recommender system.

Recommender systems are based on several techniques. The first type of system is content based, in which recommendations are based on items that are similar to those used in the past. Another type of recommender system is based on the demographic profile of the user. A third kind deploys specific domain knowledge: what aspects of items are the most useful in a particular environment? Community based systems use recommendations of the user's friends. Finally, hybrid systems combine several of the discussed techniques. The common factor is creating an extensive profile of users at the level of the individual not limited to their personal preferences, and including data about their peers. Furthermore, this profile is updated over time to keep abreast of changing preferences. From a privacy point of view, this leads to the question of trust: how much personal information should such a system contain?

Trust in recommender systems has been investigated by Chellappa & Sin (2005), from a slightly different angle: under what conditions are people willing to allow vendors to store personal information? They conclude that people are prepared to share information if the vendor is able to invoke trust. The level of trust invoked by a specific vendor is a reason for consumers to shop there, and ignore others with virtually the same offering. Even while people feel a general concern about sharing private data in general, they might be willing to give up some of their privacy in return for benefits provided by the vendor.

Not everybody will be trading privacy for convenience, and Jeckmans *et al.* (2013) have investigated possible remedies, such as raising awareness about privacy issues and invoking specific laws dealing with personal information. These types of measures have serious drawbacks. As we have seen, being aware does not stop people to engage with recommender systems and most legislation will take quite some time before coming into effect. The authors also describe technical measures such as anonymization, randomization and the use of cryptography. If user data is anonymized, the identifying information is removed, while preserving the rest. Randomization and differential privacy techniques aim to make the data of a specific person indistinguishable from most other users, by adding random data. Cryptography is considered to be a more secure choice, but with additional costs: it requires extra resources and may slow down the system.

These techniques add extra complexity to the system. This raises a question for the system's owner that mirrors the privacy trade-off by customers. Improved privacy protection will most likely have a negative effect on the system's efficiency, reducing the likelihood of implementation.

6.2.2 Libraries, privacy and the role of the catalogue

Global library cooperative OCLC lists at the time of writing 139 web based collections of open access documents (OCLC, 2016). All these collections fall within the definitions of digital libraries as discussed by Borgman (1999): a combination of "content collected on behalf of user communities", which also functions as an "institution or service". So, when a digital library collects and maintains a collection of documents in order to serve the information needs of specific groups of users, it functions as a 'traditional' library.

If open access libraries share traits with traditional libraries, we might also expect the same attitude towards privacy. The privacy of library patrons must be protected, including user data collected in library systems. This position is shared among the International Federation of Library Associations and Institutions (2016), the American Library Association (2014) and several other national library associations..

Protecting library patron's privacy is not an easy task. American libraries struggle with the implication of the USA PATRIOT Act, which expand the abilities of law enforcement agencies to collect personal information (Jaeger, McClure, Bertot, & Snead, 2004). The gathering of this type of data is not limited to the United States, but is also becoming more common in European countries (Nijboer, 2004). Apart from governmental organisations, libraries might also develop policies about other third parties who might be interested in the data generated by – and about – users (Corrado, 2007). Some libraries try to resolve the trade-off between extra functionalities and better service versus protecting personal information by adding recommender functions to their online public access catalogue (OPAC), based on anonymised usage data (Geyer-Schulz, Neumann, & Thede, 2003; Mönnich & Spiering, 2008).

Whether a 'anonymized' OPAC is a fitting solution for open access libraries can be called into question. Firstly, the role of the catalogue as primary entrance to the collection is being re-evaluated, as illustrated by Dempsey (2006). He argues that library catalogues are too limited as tools to discover content. This is put into practice at the library of Utrecht University, through the deprecation of their OPAC system. Instead, relevant results must come from search engines and library aggregators (Kortekaas & Kramer, 2014). Others are discussing whether social media websites such as Facebook.com offer an alternative. Scale (2008) concludes that Facebook does not deliver optimal results, but the article's number of citations indicates the interest in the library community. Secondly, compared to 'traditional' libraries, open access libraries – which are by definition online – might even be more depending on search engines or other external discovery tools. This is illustrated by the OAPEN Library. Its website functions as an OPAC; however, over 70% of its usage bypasses the website. The documents are accessed by enabling integration into the user's systems, the infrastructure used on a daily basis (Snijder, 2014a).

In short, privacy should be a concern for open access libraries and the OPAC – even when it does not retain reader data – might not be the best solution for content discovery.

6.2.3 Clustering books and readers through social network analysis?

The previous sections made clear that recommendation systems only function well at the cost of privacy. In a library context, this is not acceptable and the offered solutions are not ideal, especially in the case of open access libraries. This leads to the central question to discuss in this paper: how to support library users in an environment that minimizes the amount of information stored about individuals?

When it is not feasible to create profiles of individuals, we might look at the combined behaviour of all users of the digital library. Are all books downloaded at random, or can we discern clusters of books that are meaningful for groups of readers? The clustered books can be seen as a network of interconnected objects. If it is possible to identify such networks, it might be possible to recommend relevant books based on usage patterns. We might go a step further, and examine if the groups of individuals connected to those book clusters share a common trait. Thus, we are examining possible networks of books and readers.

How can we study such networks? Open access libraries do not register individual users, but a small amount of – publicly available – information about the internet provider can be used. Typically, the usage amounts to thousands of document downloads, where the provider and the document can be linked. In other words, the provider acts as a proxy for the reader. A certain document can be linked to multiple providers and one provider may be connected to multiple documents. These kinds of relations are studied using social network analysis techniques. Using graph – or network – theory, the characteristics of networks can be described and examined. Which aspects of the nodes – the parts – and the edges – the relations between the nodes – are most relevant depend on the characteristics of the network.

The possible combinations of providers and titles are quite large and thus, finding meaningful clusters is not easy. The same problem – at much larger scale – can also be found on the web. Kumar, *et al.* (1999) deployed graph theory to find "implicitly defined communities" using sets of interlinked Web pages. They aimed to find groups of content creators sharing a common interest. According to the authors, those groups could provide valuable information resources for interested users, uncover some of the sociology of the Web and target advertising. This aligns with the role of online libraries: providing valuable information to interested parties and directing them to the right documents is a core task. Finding communities in digital libraries is the first step to recommending useful content; not to individuals but to groups.

The extensive introduction into social network analysis by Wasserman & Faust (1994) can be used to define the type of network under examination. In this case, the network consists of two types of groups or modes: providers and documents. These types of networks are called two mode networks. Furthermore, the relation between the providers and the books is not reciprocated: providers act on books, but – for the purpose of this paper – books are not acting on the providers. Consequently, this two-mode network is directed.

Moreover, networks consisting of actors and passive elements such as social events – or in this case: documents – are called an affiliation or membership network. Here, the analysis is based on affiliations of actors to the passive elements, on the relation between the passive elements and the actors, or on both modes simultaneously. One possible analysis of the latter kind is finding cohesive subsets of actors and passive elements. In this case, clusters of providers and books.

The solution to the problem of finding communities in networks is described by Newman and Girvan (2004). By repeatedly using an algorithm that removes edges that acts as a 'bridge' between others, all the nodes are divided into closely connected groups. Wakita & Tsurumi (2007) have created an updated version, which is used for the research in this paper.

The use of social network analysis or clustering algorithms is not limited to the discovery of user groups. For instance, Verleysen & Weeren (2016) used a "fuzzy cluster analysis" to examine the divide between authors publishing in international journals in English, compared to those writing books and chapters in national or regional languages. Their results are supported by computer generated outputs. In contrast, Provan, *et al.* (2005) encourage community leaders to use social network analysis procedures to manually describe the networks they participate in. These approaches demonstrate the breadth of social network analysis.

The procedure outlined in this paper should be relevant to all kinds of digital open access libraries, leading to some additional requirements. First, it must be applicable to a wide variety of collections. Therefore, the metadata used has to be attainable from different types of documents. The metadata used will be discussed further in section *6.3.2 The books*. Secondly, the tools to be used ought to be available as open source software, preferably with an easy to use interface. This paper's analysis has been conducted using NodeXL, a free and open source network analysis tool using a Microsoft Excel template. It is maintained by the Social Media Research Foundation, co-founded by Marc Smith (Shneiderman & Dunne, 2013).

6.3 Quantifying the data set

The previous section discussed the tension between privacy and optimizing recommendation systems. Using social network analysis to find communities around certain books might enable open access libraries to create recommendations, while retaining the privacy of the individual readers. In order to test this idea, the usage of the OAPEN library will be analysed.

6.3.1 The collection

The OAPEN Library is managed by the OAPEN Foundation, a not-for-profit organisation based in the Netherlands. The Foundation's goal is to promote open access book publishing, through building and disseminating a quality-controlled collection of open access books (OAPEN Foundation, 2016). The books discuss a broad range of subjects, and are written in several languages. Around half the publications are written in English; both Dutch and German amount to roughly 20% and 5% of the books are Italian. Other languages include French, Danish, Spanish and Latin. Section *6.3.2 The books* describes the state of the collection in 2012.

As stated before, our goal is to find clusters of books and providers that have a meaningful connection. In other words, we need to establish whether a combination can be attributed to an underlying theme, and not determined by chance. This procedure must be transparent and reproducible in other collections than the one currently under examination. The method used is based on quantifying several aspects of the total collection. These numbers are compared to the amounts measured in the clusters.

The data describing the complete collection of 2012 and 2014 is available in the appendix. It will be used as benchmark. The data of the clusters described in section 6.4.1 and section 6.4.3 is also listed there. The appendix, the underlying lists of downloads, providers and the clusters they occupy are available via http://dx.doi.org/10.17026/dans-x72-d9h2.

6.3.2 The books

The clusters contain books and providers; the first step is to determine which aspects will be examined. Starting with books, the number of possible aspects is large. The books are collected and maintained by the owner of the digital library, who might choose to describe the documents in many different ways. A typical book description in a library catalogue contains the title, author, publisher, place and year of publication, number of pages, ISBN, language, whether it is part of a series, and indications of the book's subject through keywords and classification codes. However, these descriptions serve several purposes: some are useful to identify a work, while others may help to indicate the book's topic and its quality or prestige. In this case, we assume that the users of the electronic library are interested in books "about" a subject.

In general, the contents of a scholarly book will not be limited to one subject. Even if the authors are exploring one theme, the book will discuss several facets. An example is the book "Malaysian Cinema, Asian Film: Border Crossings and National Culture".[1] This book might be useful for those who are interested in film and media studies, but also for those who are involved in the culture of Malaysia or Southeast Asia.

[1] See http://oapen.org/search?identifier=340243.

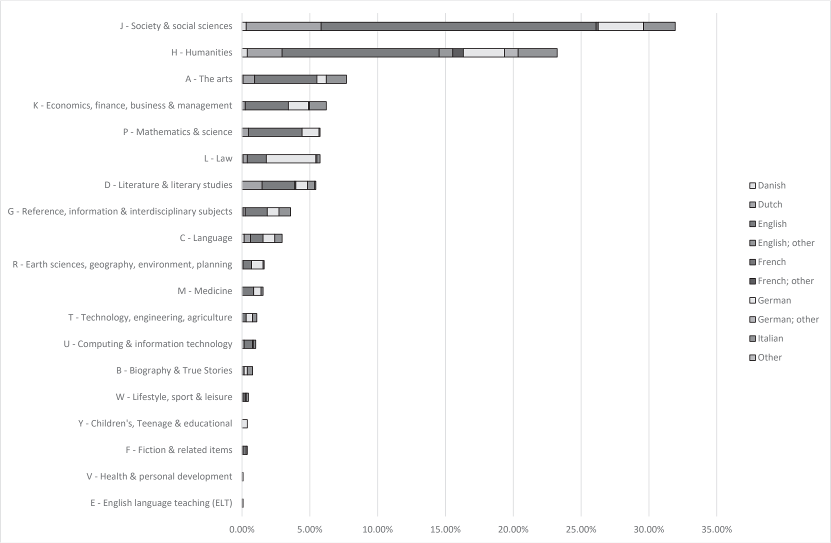
Another aspect to consider in a web-based collection is language. Open access libraries are open to everyone with an internet collection; a potential worldwide audience. However, it is unlikely that prospective readers will download books in languages they cannot understand.

The most widely used methods to describe the contents of document are keywords and classifications. Keywords are potentially more accurate, but the endless possibilities make it hard to create quantifiable sets. In contrast, classifications are based on a hierarchy. By selecting the top levels, it is possible to create relatively few sets of books.

Compared to keywords, the possible number of languages is lower. A large variation in languages is not always useful in an analysis. For instance, there is no point in listing Swedish as a separate language if a collection of thousands of titles contains three books in Swedish. Thus, the quantification is based on a simplification of the available metadata.

For the analysis, each book in the collection was categorised as follows: it belongs to one language set, and it may contain up to five subject codes. These two aspects can be placed in a matrix, serving as a ‘snapshot’ of the collection in a certain point of time. Such a matrix is quite useful in displaying every possible detail, and helps to quantify the amount of titles in a certain language or subject. Still, a visualisation is a more optimal way to display the main characteristics of the collection. For instance, the large percentage of documents on society and social sciences or humanities and the amount of English language books are easily spotted in *Figure 1*. Other details – such as the large percentage of German language books on law – are also visible.

Figure 1 OAPEN Library collection: languages and subjects (2012)



6.3.3 The providers

In contrast to books, the amount of information about providers is limited. When users are not tracked, there is little more available than the name of the provider and the time when a specific book was downloaded. Identifying individual readers is next to impossible: a feature that protects the reader's privacy. Using public information – based on the WHOIS internet protocol (Daigle, 2004) – the provider's country of origin is also available. Thus, using the provider as proxy, readers can be grouped by nation. This method is not 100% accurate: when a Dutch native travels through Canada and accesses a book using a Canadian internet provider, this will be listed as a "Canadian" download.

Logging internet providers leads to another interesting question: how many people have downloaded books through that provider, and are they interested in the same things? If the provider is an organisation with a strict focus, chances are that all members share a similar interest. An example is the organisation Bouwkennis – a Dutch marketing firm, specializing on the building sector, which downloaded a report on housing policy. In contrast, also listed among the downloading companies is Verizon, a large internet service provider serving millions of customers. It is highly unlikely that all documents downloaded through Verizon are the result of a single person or a 'single minded' group. Another complication is the freedom of users of an online open access library to download as many titles as they like.

The question is how to select download patterns that are the result of a single person's action, or the action of a goal-oriented organisation. The number of titles downloaded by a provider should not be the deciding factor. On the contrary, we might imagine several readers who are interested in the same twelve books: the kind of pattern that hints at a shared interest. On the other hand, we need to filter out the actions of a diffuse group of people who only happen to share the same internet provider. The solution chosen here is to look at the number of times a single title is downloaded through a provider. The number of downloads are logged per month. To be absolutely sure that a single person has downloaded a title, multiple downloads of the same title by the same provider in one month are discarded. All downloads where just one copy of different titles is downloaded by a single provider are still part of the data. Besides, if the provider downloads the same title in another month, this download will also be part of the analysed data.

How does this choice effect the data? The download data for 2012 – collected during three months – consists of 6,176 providers who downloaded 57,508 books. After removing those providers that have downloaded the same title more than once in the same month, the number of providers becomes 5,180 (84% of 6,176) and the number of books downloaded is 34,345 (60% of 57,508). The majority (53%) of the 5,180 providers downloaded a single title; amounting to 2,740 providers. The remaining 47% (2,440 providers) downloaded between two and 338 different titles. Examining the number of providers that download more than one book demonstrates that the majority of that group (1,440 providers) never 'take' more than five books. This is consistent with the assumption that we are looking at individuals that search for specific titles, instead of those who are downloading as much as possible.

The ratio of nationalities is next to be examined. The percentages of all visitors of the online library can be used as a benchmark to compare against the clusters. A cluster containing a considerable difference in nationalities combined with a substantial difference in the range of subjects might signal that the books and providers have a meaningful connection. To enable this, we need to list the nationalities of all providers. However, the data contains over 160 different countries, ranging from Albania to Zimbabwe. The goal is to find significant differences, not a complete list. Therefore, the benchmark can be simplified to the ten countries with the highest usage. When the 'top-10' of a cluster contains countries not in listed in the benchmark, this is a clear – and easy to detect – signal.

6.3.4 The influence of the collection

The previous sections discussed the choices made to quantify the most useful aspects of the publications in the collection, and its users. These aspects are used to analyse how the readers – through the providers – interact with the books in the digital library. In other words: the collection shapes the possible actions of the readers. This leads to the question whether changes in the collection lead to changes in usage. To test this, the same investigation is carried out using data from 2014, two years after the first analysis. During that period, the collection of the OAPEN Library doubled from just over 1,100 titles to more than 2,300 titles. This growth influenced the collection on both axes: subject and language.

The growth of the collection altered the ranking of the subject categories and the languages. In 2012 "A - The arts" ranked third and "K - Economics, finance, business & management" ranked fourth. In 2014, this was reversed. The same holds true for "P - Mathematics & science" – ranked sixth in 2012 and seventh in 2014 – and "D - Literature & literary studies" – ranked seventh in 2012 and sixth in 2014. Within the languages, the ranking of Danish changed from fifth to seventh. Furthermore, due to the influx of titles in English, German and Dutch, the percentage of Italian language titles plummeted from 11% in 2012 to 5% in 2014.

The differences between 2012 and 2014 indicate that the focus of the collection may have shifted. Does this also lead to differences in usage?

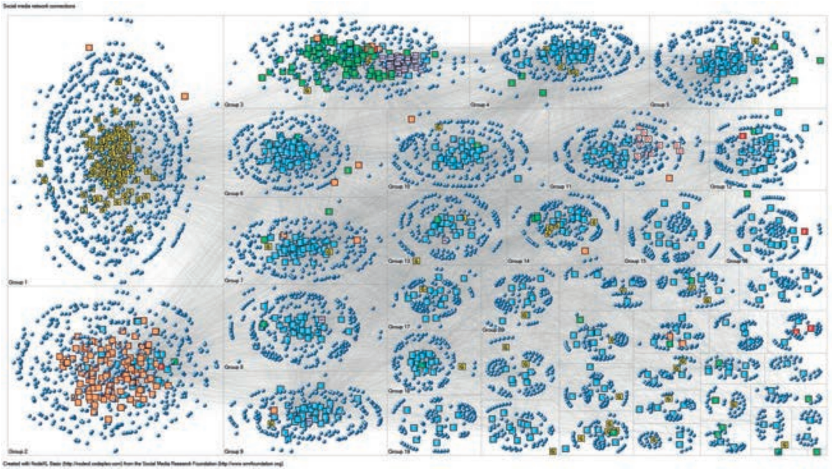
6.4 Analysis

The previous section discussed the way the book and provider data was quantified: what aspects are to be examined? Furthermore, the collection of the OAPEN Library has expanded extensively, which also affects the usage. In this section, the analysis of the collection's usage in 2012 and 2014 are presented, exposing the difference in download patterns.

6.4.1 Examining clusters – the OAPEN collection in 2012

The first step towards answering the questions discussed in the previous sections is examining the usage patterns that occurred in 2012. In section 6.4.3, book downloads in 2014 are examined and the differences will be discussed. During the examined period – lasting three months – 967 different titles were downloaded by 5,180 providers. The total number of downloads is 34,345.

Figure 2 Clustered providers and books in the OAPEN Library, 2012


Social media network connections
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
Group 7
Group 8
Group 9
Group 10
Group 11
Group 13
Group 14
Group 15
Group 16
Group 17
Created with NodeXL Basic (http://nodexl.codeplex.com) from the Social Media Research Foundation (http://www.smrfoundation.org)

The linked titles and providers are clustered using the Wakita-Tsurumi (2007) algorithm, resulting in 43 clusters ranging in size from 1,000 elements (125 books and 875 providers) to a cluster consisting of exactly one book and one provider. And so we need to consider the number of clusters to investigate. At what point is the cluster too small to convey meaningful results? As this kind of study is scarce, there are no tried and tested guidelines. This paper's result might be considered to be a proof of concept, where the additional question of the optimal number of clusters is ignored for now. Instead, the – somewhat arbitrary – boundary is set at the ten largest clusters.

It must be noted that the data used in the Wakita-Tsurumi algorithm consists of nothing more than a unique code for each book, and the name of the provider. For instance, the connections between "uni-mannheim.de" and "422333"; "uni-mannheim.de" and "391039" are part of the data. After deploying the algorithm, the provider "uni-mannheim.de" is classified as German, and the books are identified as *Vernetztes Leben. Soziale und digitale Kulturen* and *The Practices of Happiness : Political Economy, Religion and Wellbeing*. Thus, the algorithm cannot be influenced by aspects of the providers or the books.

The data of each cluster has been analysed based on the following procedure. Firstly, the ranked subject classifications and the languages of the books in each cluster are compared to the complete collections' data. The second question is whether the cluster's providers division in nationalities are in line with the percentages for the complete set. Substantial changes trigger a further examination of the book's subject by assessing keywords and titles.

The analysis resulted in the following 'named clusters':

- Cluster 1. German language books. Books in the German language, mostly downloaded by readers from Germany, Austria and Switzerland.
- Cluster 2. Dutch language books. Books in the Dutch language, mostly downloaded by readers from The Netherlands and Belgium.
- Cluster 3. Italian language books. A majority of the books is written in the Italian language, mostly downloaded by readers from Italy.
- Cluster 5. Film and Media. Books in the English language. The cluster contains a large group of books discussing film studies, plus a few titles on media or theatre studies.
- Cluster 6. Migration. Books in the English language, focused on migration studies.
- Cluster 9. Indonesia and South-East Asia. Books in the English language, mostly discussing Indonesia, in combination with works on South East Asia.

The appendix and the complete data set are available here: http://dx.doi.org/10.17026/dans-x72-d9h2.

6.4.2 Analysis results – 2012

Table 1 OAPEN Library: cluster analysis results (2012)

Cluster	Title	Books: Subject classifications	Books: Language	Readers: Nationality	Books: Keywords
1	German Language books	L – Law ranked #1, compared to #6. Most books on Law are written in German.	97% German, compared to 21% in the data set as a whole.	Germany ranked #1 (65%), #2 Austria (10%), #3 Switzerland (8%). Both Austria and Switzerland are not part of total collection top 10.	
2	Dutch Language books	D - Literature & literary studies ranked #3 compared to #7.	81% Dutch, compared to 11% in the data set as a whole.	Netherlands ranked #1 (64%), Belgium ranked #2 (12%). Belgium is not part of the total collection top 10.	

Cluster	Title	Books: Subject classifications	Books: Language	Readers: Nationality	Books: Keywords
3	Italian language books	G - Reference, information & interdisciplinary subjects ranked #3, compared to #8.	64% Italian, compared to 10% in the data set as a whole.	Italy ranked #1 (44%).	
5	Film and Media	A – The arts ranked #1, compared to #3.	92% English, compared to 52% in the data set as a whole.	USA ranked #1 (42%), #2 Great Britain (13%).	27 of the 40 titles discuss film studies.
6	Migration	J - Society & social sciences ranked #1 (64 % compared to 31 %).	88% English, compared to 52% in the data set as a whole.	USA ranked #1 (22%), #2 France (13%) compared to #6, #3 Spain (12%). Spain is not part of the total collection top 10.	38 of the 47 titles discuss migration.
9	Indonesia and South-East Asia	J - Society & social sciences ranked #1 (45 % compared to 31 %).	92% English, compared to 52% in the data set as a whole.	USA ranked #1 (32%), #2 Indonesia (18%), #3 Australia (9%) compared to #10. Indonesia is not part of the total collection top 10.	22 of the 35 titles discuss Indonesia or South-East Asia.

The role of nationality and language is visible in the largest clusters. The first cluster contains 125 books, and 122 of those titles are in the German language. Also, the 'top 3' nationalities of the providers are German speaking countries. The same holds true for the second cluster, which consists of a large majority of Dutch language monographs. Here, the Dutch and Belgian providers rank one and two, respectively. And the third cluster is dominated by Italian languages monographs and Italian providers. Within these clusters, the ranking of the subjects seems to reflect the division of the books in the respective language. For instance, in Cluster 1, Law ranks first. The explanation can be found in the relative large number of law titles in the German language.

Within the cluster on Film and Media, English plays a major role. As is the case with Cluster 6 and Cluster 9, the USA providers are now ranked at the first place. In these three clusters, subject plays a major role. This can be inferred from the differences between the classification within the clusters compared to the whole collection, and an examination of the keywords and titles of the books. We might conclude that the role of English is different from German, Dutch or Italian: it is not a defining property of a cluster.

In the case of Cluster 9 – Indonesia and South-East Asia – the interest through Indonesian and Austrian providers can easily be explained by a regional focus. In contrast, the international usage of the clusters on Film and Media or Migration do not show such a clear pattern. In the case of Cluster 6 – Migration – the spread of providers is relatively even: there are no countries with a much stronger interest compared to the other 'members' of the cluster. It is noteworthy however, that the "providers top 10" lists Spain, Greece, Austria and Hungary. All of these countries are not part of the total collection's top 10. Yet, in these countries – and also in France, Poland and Germany – immigration is a widely-debated topic. This might point to a regional interest, but the signal is not as strong compared to the data by Cluster 9. See the Appendix for more details.

6.4.3 Examining clusters – the OAPEN collection in 2014

When the same method is applied to the data of a three-month period in 2014, the differences are striking. During that time, 2,334 different titles were downloaded 60,238 times, roughly twice the amount of 2012. However, the number of providers 'only' raised 20% to 6,316 providers. Most of these providers (69%) downloaded one book in a month; and the total percentage of providers that downloaded 5 titles or less is 98%. Furthermore, the 'country top ten' list contains the same countries, with the exception of Ukraine, which replaces Poland.

The question is whether the changes in the collection affected the usage: is it possible to detect the same clusters? Here, the number of clusters is comparable to 2012: 41. The largest cluster contains 244 books and 723 providers, while the smallest cluster consists of one book and two providers. Again, the largest ten clusters are compared to the data of the complete collection.

The analysis resulted in the following 'named clusters':

- Cluster 1. German Language books. Books in the German language, mostly downloaded by readers from Germany. Providers from Austria and Switzerland are ranked third and fourth. Ranked #2 are providers from the US.
- Cluster 2. Dutch Language books. Books in the Dutch language, mostly downloaded by readers from The Netherlands. Comparable to Cluster 1, the US providers rank second, followed by Belgium providers.

- Cluster 3. The largest clustering of Italian language books, but this cluster also contains a large portion of German books. Here, Italian rank first, followed by German providers.
- Cluster 5. Indonesia and South-East Asia. This cluster is comparable to Cluster 9 of the 2012 data, containing books in the English language, mostly discussing Indonesia and South-East Asia.
- Cluster 9. Australia and the Pacific region. English language books on subjects related to Australia and the Pacific region. The US providers are ranked first, Australian second.

6.4.4 Analysis results – 2014

Table 2 OAPEN Library: analysis results (2014)

Cluster	Title	Books: Subject classifications	Books: Language	Readers: Nationality	Books: Keywords
1	German Language books	R - Earth sciences, geography, environment, planning #4, compared to #10. Most books on this subject are written in German.	91% German, compared to 24% in the data set as a whole.	Germany ranked #1 (46%), #3 Austria (10%), #4 Switzerland (6%). Both Austria and Switzerland are not part of total collection top 10.	
2	Dutch Language books	Consistent to the number of titles in Dutch, K - Economics, finance, business & management#2 and H - Humanities #3	92% Dutch, compared to 19% in the data set as a whole.	Netherlands ranked #1 (52%), Belgium ranked #3 (9%). Belgium is not part of the total collection top 10.	
3	Italian language books		23% Italian, compared to 5% in the data set as a whole. However, 43% of the books are German.	Italy ranked #1 (22%), #2 Germany (20%).	

Cluster	Title	Books: Subject classifications	Books: Language	Readers: Nationality	Books: Keywords
5	Indonesia and South-East Asia	J - Society & social sciences ranked #1 (43 % compared to 31 %). K - Economics, finance, business & management ranked #2 (16% compared to 8%)	98% English, compared to 47% in the data set as a whole.	Indonesia ranked #4, #5 India, #6 Pakistan. Indonesia, India nor Pakistan are part of the total collection top 10	60 of the 119 titles discuss Indonesia or South-East Asia.
9	Australia and the Pacific region		70% English, compared to 47% in the data set as a whole.	Australia ranked #2 (22%) compared to #10.	74 of the 124 titles discuss Australia or the Pacific region.

The three largest clusters are once again connected to books in a specific language, without a specific emphasis on a subject. We could argue that the contents of Cluster 3 are relatively 'diluted': the number of German books is higher than the books in Italian. However, it contains the largest concentration of Italian monographs, combined with a large Italian readership.

It is noteworthy that both Film studies and Immigration are less visible, while books focusing on the Oceania region are easily spotted. An explanation may be found in the influx of new titles in the OAPEN Library. In 2013, the collection grew with over 300 titles published by ANU Press, part of Australian National University.

The number of titles on immigration did not grow as spectacular. Snijder (2013) discusses the dissemination of books by the IMISCOE Research Network on international migration, integration and social cohesion. The majority of those book made available through the OAPEN library in 2012, and the data of 2012 contains 50 IMISCOE titles. Most of them – 34 books – are found in cluster 6: Immigration. Between 2012 and 2014, only ten more titles were added – a total of 60 books. Compared to the growth of the complete OAPEN Library collection, this is a modest increase.

The role of American providers is also unmistakable. According to *The World Factbook* (Central Intelligence Agency, n.d.), the number of American internet hosts is 505,000,000. A large number, compared to the second country on the list – Japan – which contains 64,453,000 hosts; a factor 7 less. Given these amounts, the prominent role of US providers is not surprising.

6.5 Creating recommendations based on clusters

We have discussed before that personal recommendation systems cannot be used in open access libraries. It is nonetheless possible to detect patterns in the use of the library, and with relative simple means, meaningful clusters of books and providers can be detected; the current results can be seen as a proof of concept. Contrary to the assertion of Lynch (2002), it is possible to identify – up to a point – which user communities will engage with the digital library. The detected patterns help at the very least to define interests by larger groups of readers; a precondition for the creation of new services.

A possible service could entail listing groups of titles, to be presented to certain groups of providers. The clustering results can be converted into a set of recommendation rules, based on the contents of the book combined with the nationality of the provider. For instance, the results from cluster 1 of section *6.4.2 Analysis results – 2012* could be transformed into the following 'recommendation rule': If the provider is based in Germany, Austria or Switzerland and has downloaded a book in German, present a list of all German books. Likewise, this combination of provider nationality and non-English books could be applied to cluster 2 – Dutch language books – and cluster 3 – Italian language books.

There are also subject-based clusters, for instance cluster 5 in section *6.4.4 Analysis results – 2014*. Here, the recommendation might run along the lines of presenting English language books on Indonesia or South-East Asia to providers based in Indonesia, India or Pakistan. Cluster 5 of section *6.4.2 Analysis results – 2012* would lead to a more generic rule: if one English language book on film and media studies has been downloaded, present all English language titles on this subject.

The suggestions listed are not the result of a careful curation by a librarian, but are purely based on the usage patterns that have been uncovered. Recommender systems are based on the preferences of individuals; the suggestions here are based on the preferences of "implicitly defined communities" as described by Kumar *et al.* (1999). In this way, the flexibility of recommender systems is deployed, without violating the privacy of individuals.

6.6 Discussion

In the previous sections, we have seen the analysis and the possible recommendations based on its results. Yet, on a more abstract level there are several other aspects to reflect upon: the role of regional interests and how well the deployed algorithm performs on the total collection.

When the different clusters are analysed, the influence of language communities is profound. It might not come as a surprise that readers in languages other than English tend to be more interested in publications written in their 'local' language. Thus, the clusters of books in German, Dutch or Italian are read mostly by native speakers. The language effect is quite strong: within these clusters it is hard to find a subject based focus. In contrast, if publications in English are taken into account it is still possible to find clusters whose subject is closely tied to a region. This is especially visible in the clusters focused on Indonesia and Sea East Asia, and the cluster concerning Australia and the Pacific region. Even the subject of immigration could be seen as a regional – mostly European – concern.

One might argue that the available data tends to point in this direction: the main thing known about the readers are their provider's countries. Furthermore, one of the aspects analysed is the distribution of nationalities. Given this procedure, it is rather hard to miss 'regional' patterns. On the other hand, region is not the only scrutinised aspect. The books' subject and language are also taken into consideration. As an additional test, all twenty clusters have been analysed using subject and language only. This did not lead to new insights.

The clusters are the results of algorithms – predefined procedures. Deploying these procedures leads to interesting results: uncovering usage patterns. We also saw that the results differ: in 2014, no clusters concerning Film and media or Immigration were detected. Instead, the large influx of books by an Australian publisher was visible. Does this mean that interest in film or immigration studies has diminished? That might be possible, but another option seems more likely: the performance of the algorithm when it is applied to the collection.

Between 2012 and 2014, the collection nearly doubled. As was illustrated by the example of the IMISCOE series and the books in Italian, the number of books concerning a certain subject may not always keep pace with the collection's growth. The algorithm only detects the 'strongest' patterns, based on relatively large groups. Thus, smaller clusters of books and readers may go unnoticed.

The background section discussed several types of recommender systems. The variety hints at room for improvement: there is no single best solution. This may also apply to this paper's procedure; other procedures and algorithms may yield improved results. A recent paper by Gläser, Glänzel, & Scharnhorst (2017) illustrates this: the authors describe the search for optimized deployment of algorithms to cluster articles into “thematic clusters”. Different algorithms lead to different results, all of which might be valid in their own way. The theme of this paper is also a clustering problem, and thus the results by Gläser *et al.* could be applied here as well.

Simply put: the question is how to proceed from this starting point in order to create a fully functioning system? There are several points to explore. Firstly, the results of several clustering algorithms should be evaluated. We have seen that the currently used algorithm detected other groups in the collection data of 2012 and 2014. Will other algorithms lead to strongly differing results? Another avenue to explore is recursive use: deploying the algorithm again on the clusters, in order to find ‘sub groups’. Earlier in the paper, the question which clusters should be investigated was mentioned. This might be an additional study. Lastly, the current analysis is depending on human judgment, especially on the book’s subjects. In an open access library, the documents are available in a full text form. Using text mining techniques might help to automatically cluster the books, based on common words or word sequences. It would be interesting to see if these ‘subject clusters’ overlap with the clusters of providers.

6.7 Conclusion

This paper attempts to unravel the paradox of open access libraries: created for maximum dissemination, but deploying one of the most powerful tools to support its users leads to questions about privacy. Recommender systems are used widely and with great success, but are built on storing information about individuals. This is hard to accept from a privacy point of view, and open access libraries are not normally equipped to individually track their readers. However, every library functions better when it understands the needs of its patrons.

Open access libraries are web based by definition, and the usage through providers indicates the level of interest for each document. The thousands of data points require the use of automated procedures. Applying social networking analysis techniques helps to uncover patterns of usage that are very hard to spot in a different way. With relative ease, it is possible to run a meaningful analysis of the interests of groups of readers.

This paper’s results can be seen as a proof of concept; a possible starting point for recommendations built on usage that retain the privacy of individual readers.

6.8 Acknowledgements

The author would like to thank Professor Paul Wouters of the Centre for Science and Technology Studies (CWTS) and Professor Frank Huysmans of University of Amsterdam for commenting on the draft version of this paper.