Received: 28 March 2023 | Revised: 16 October 2023 | Accepted: 6 March 2024

DOI: 10.1002/asi.24886

RESEARCH ARTICLE

JASIST WILEY

Measuring the retrievability of digital library content using analytics data

Hamed Jahani[1] | Leif Azzopardi[2] | Mark Sanderson[3]

[1]School of Accounting, Information Systems and Supply Chain, RMIT University, Melbourne, Victoria, Australia

[2]University of Strathclyde, Glasgow, UK

[3]School of Computing Technologies, RMIT University, Melbourne, Victoria, Australia

Correspondence
Mark Sanderson, School of Computing Technologies, RMIT University, Melbourne, VIC, Australia.
Email: mark.sanderson@rmit.edu.au

Funding information
Australian Research Council Linkage Infrastructure, Equipment and Facilities (LIEF) Program, Grant/Award Number: LE180100094

Abstract

Digital libraries aim to provide value to users by housing content that is accessible and searchable. Often such access is afforded through external web search engines. In this article, we measure how easily digital library content can be retrieved (i.e., how retrievable) through a well-known search engine (Google) using its analytics platforms. Using two measures of document retrievability, we contrast our results with simulation-based studies that employed synthetic query sets. We determine that estimating the retrievability of content given a Digital Library index is not a strong predictor of how retrievable the content is in practice (via external search engines). Retrievability established the notion that search algorithms can be biased. In our work, we find that while there such bias is present, much of the variation in retrievability appears to be strongly influenced by the queries submitted to the library, a side of retrievability less examined in past work.

1 | INTRODUCTION

Critical to owners of a Digital Library (DL) is an understanding of how their content is used. Such information can provide guidance on how the DL's interface and system components can be improved or optimized; determine what DL content is providing value; and/or guide future acquisitions for the DL's collection. In a review of past such analyses, Kelly (2014) showed that most examinations of DLs employed methodologies such as user studies or the interrogation of transaction logs, often via Google Analytics (GA). While of great value, there is an aspect such methodologies do not examine: the *retrievability* of DL content. Azzopardi and Vinay (2008c) were the first to study retrievability, defining it as a measure of how likely a document is to be retrieved given the search engine or search algorithm used. In a series of articles—Azzopardi and Vinay (2008a, 2008b); Azzopardi and Bache (2010); Wilkie and Azzopardi (2013, 2014, 2016, 2017, 2018)—Azzopardi and his collaborators showed that the collection documents indexed by a search engine are not equally retrievable. The search algorithm employed plays a key role in deciding which documents are more or less retrievable.

However, there is a gap in the existing literature: the vast majority of retrievability work was tested on information retrieval (IR) offline collections and not on working DLs. Past research also assumed that retrieval would take place on the internal search engine of a DL. However, the administrators of most DLs allow the content of their library to be indexed by web search engines, such as Google, Bing, Baidu, and so on. Understanding the retrievability of DL content through such external services would provide new insights into what is being retrieved and how likely particular types of content will be retrieved. Retrievability research has not examined this aspect of a library's operation. Prior work aimed to assess what could and could not be retrieved to determine a retrieval algorithm's biases. In such past work, test queries were simulated, words were sampled from the documents of the DL collection. However, to obtain insights into what is actually being retrieved, real queries submitted to a DL should be used instead. Exploiting a relatively recent tool (the Google Search Console, GSC), this article takes the method proposed by Azzopardi and Vinay (2008b) and applies it to a DL, measuring the retrievability of that library's content based on the queries submitted to a large external web search engine, Google. We calculate the retrievability of documents from the DL as searched by Google using a query set recorded by GA.

Our work allows us to answer the following research questions:

1. What is the retrievability of the documents in a DL to the users of an external web search engine?
2. Can we predict retrievability scores of documents in the DL based on the simulation methods detailed in past work?
3. Can we identify which features correlate with higher retrievability in the external web search engine?

The rest of this article reviews the existing work conducted in retrievability, followed by a description of the methodology and the data used. The results of the retrievability analysis is detailed next, followed by a comparison of that work with past retrievability methodologies. Finally the conclusions and future work are outlined.

2 | LITERATURE REVIEW

There is a long tradition of research on the evaluation of IR systems, with work starting in the 1950s, see Sanderson (2010) for a historical perspective. That work focused on measuring the ability of retrieval systems to return documents that were relevant to a user's query. Another strand of evaluation research emerged as online searching systems became more prominent. Here, the assessment of the way that users interact with a live retrieval system was examined, measuring clicks and user engagement with retrieved documents. From that interaction data, inferences were made on the effectiveness of searching systems (see Hofmann et al., 2016; Yue et al., 2010). While researchers strove to ensure biases did not affect the focus of their measurements in either of the evaluation modes, one bias not considered in these works was the question of whether the documents of a collection were equally retrievable. Two aspects of this question have been addressed: personalization and *retrievability*, we focus here on retrievability.[1]

Azzorpardi hypothesized that a retrieval algorithm may contain inherent biases that could cause some documents in a collection to be ranked higher than others. His hypothesis was tested by measuring the retrievability of documents indexed by a given IR system (Azzopardi & Vinay, 2008a). Across a series of experiments based on simulations, Azzopardi found that his hypothesis was supported. Azzopardi et al. examined multiple aspects of this topic (Azzopardi & Vinay, 2008a; Bache, 2011; Bashir & Rauber, 2010; Wilkie & Azzopardi, 2018). Details of that work are discussed in Section 3 of this article.

Azzopardi's work has been used in pragmatic settings: Bashir and Rauber (2010) drew queries from a log that was submitted to search for prior art in patent searching. The authors employed retrievability to understand the location of documents that were retrieved by those queries and used the analysis to show that the queries failed to retrieve a large number of relevant patents. Roy et al. (2022) used retrievability to study the retrieval biases in a DL composed of publications and also datasets that could be searched. The authors found that retrieval of datasets was more biased than retrieval of publications. Both works demonstrated the practical utility of retrievability analysis.

In this article, we consider an aspect of Azzopardi's retrievability methodology, the query set, which in past work was almost always generated by sampling words and pairs of words from the collection of documents under consideration. Across all such queries, retrievability was measured by examining how likely documents were to be retrieved and where they were ranked relative to other documents. A question that Azzopardi and colleagues did not address was how realistic was the generated query set? To explore this aspect, Traub et al. (2016) drew on the query log of a DL comparing retrievability results from the queries of the log with retrievability from a set of generated queries, which the authors referred to as "simulated." A "substantial difference" was found between the retrievability results across the two query sets. See also recent related work on "exposing queries" Li et al. (2022).

In this article, we re-examine the question of the generated query set, comparing retrievability results from the generated set and a query log. We do this comparison in the context of how an external web search engine searches DL content and how retrievability analysis can help to understand this side of a DL's operation. This aspect of DL evaluation has not been as extensively examined by the DL research community. Many of the articles that evaluate DLs focus on the library itself and not its external interaction with the wider information environment. In a modern context, DLs are websites. The core means of accessing content on the web is through popular web search engines such as Google, Bing, or Baidu. Such search engines will often crawl the content of DLs and make that content accessible to their users, however, many DL evaluation papers do not considered this aspect in their evaluation, see Fuhr et al. (2007) and Li and Liu (2019), for example. However, as was made clear in a 2010 survey of libraries De Rosa et al. (2011), while online libraries are still used, almost no one starts their search in a DL, they start with a web search engine. Examining how DL content is accessed by such engines is a crucial and somewhat overlooked aspect of DL evaluation. We wish to determine whether our internal assessment of how retrievable the content in a DL is predicts its external retrievability (and how retrievable it actually is via search engines).

3 | METHODOLOGY

We utilize two common measures that calculate the retrievability score of the documents in a DL. We also discuss how past studies generated a query set.

3.1 | Measuring retrievability

The concept of document retrievability was introduced by Azzopardi and Vinay (2008a) who calculated the retrievability metric using the notations and definitions provided in Table 1 and the following formula:

$$R(d) = \sum_{q \in Q} Pr(q) \cdot f(\delta(q,d),\theta). \tag{1}$$

Azzopardi and Vinay note that examining a ranked list of documents incurs a cost for users that grows the further down the list they look.

Azzopardi and Vinay (2008a) offer two approaches for calculating θ in the $f(\delta(q,d),\theta)$ function. First, they use a cut-off value c, which is compared with the rank of document d in any search q ($\delta(q,d)$). If $\delta(q,d) \leq c$, the function returns 1, and 0 otherwise. This measure is called the *cumulative-based* metric. Second, following the formulation used for the accessibility of lands (Hansen, 1959), a *gravity-based* metric is defined according to Equation (2), in which, β is a dampening factor that determines the position of the document in a ranked list. In the case that $\beta = 1$, the reverse rank of the document is reflected as the retrievability of the document for the specified query, which is known as a measure of document IR system performance called the "expected search length" (Cooper, 1968).

TABLE 1 Equations notations.

Notation	Description
Indices	
d	Document index in the set of all documents (D)
q	Query index in the set of all possible queries (Q)
Parameters	
$Pr(q)$	Probability of occurrence of query q
$f(\delta(q,d),\theta)$	Utility function with the cutoff value θ as a parameter
$\delta(q,d)$	Metric of the cost associated with accessing document d given the query q
$R(d)$	Retrievability score for document d
$\overline{R}$	Average of retrievability scores for document set D

$$f(\delta(q,d),\theta) = \frac{1}{(\delta(q,d))^{\beta}}. \tag{2}$$

The cumulative- and gravity-based metrics have been used in several studies. We tabulate summaries of those studies in Table 2. Generally, the studies employ the formulas for demonstrating the effect of IR algorithmic bias on retrievability—Wilkie and Azzopardi (2013, 2016, 2017, 2018); the influence of document features, such as length, on retrievability—Azzopardi and Vinay (2008b); Bache (2011); Wilkie and Azzopardi (2013, 2016, 2018); or the relationship between retrievability and IR performance metrics like precision—Azzopardi and Bache (2010); Wilkie and Azzopardi (2017).

3.2 | Corpus retrievability metrics

We detail three such metrics.

3.2.1 | Gini coefficient

Retrievability is calculated on a per document basis. However, one may also wish to have an overall measure of retrievability. One approach is to employ the *Gini* coefficient (Azzopardi & Vinay, 2008a; Bache, 2011), which was defined in economics for measuring the inequality in the distribution of income across a population (Gastwirth, 1972). The coefficient is formulated as Equation (3), computed from the retrievability ($R(d)$) over a collection with N documents. The formula G_D is zero (total equality) if all documents were equally

TABLE 2 Focus of past studies using retrievability.

Author (year)	Focus	Document set (D)		Query set (Q)		Parameters	IR model (i.e., algorithm)
		Volume	Type	Volume	Method of generation		
Azzopardi and Vinay (2008b)	Evaluating R(d) formulas for a data set and relations to the document features	NA	TREC AQUAINT collection which consists of three different news sources (APW, NYT and XIE)	663 K	Single-term most frequent words generated from documents set	$c=(20,30,\ldots,100)$, $\beta=(0.5,1.0,1.5,2.0)$	TFIDF, BM25, and Language Model (LM)
Bashir and Rauber (2010)	To increase the retrievability of patents, they expand prior-art queries generated by using query expansion with pseudorelevance feedback.	54,353	Freely available patents from the US patent	10.2 M	Two-, three-, and four-term most frequent words	$c=(30,40,60,100)$	TFIDF, BM25, Exact Match, and LM2
Azzopardi and Bache (2010)	Discussions on trade-off retrievability for effectiveness	NA	Two TREC collections: Associated Press (AP) 1988–1989 and Wall Street Journal (WSJ) 1987–1992	100 K	Two-term most frequent words derived from each corpus	$c=(10,100)$	BM25 and LM
Bache (2011)	Measuring the overall access to a patent collection and improving access to all documents by examining sensitivity to term frequency, length normalization, and convexity	19.3 M	MAREC collection (four patent offices: European, US, Japanese, and World)	2.1 M	Two-term most frequent words by using single-term words and AND and OR Boolean operators between them	$c=(10,20,50,100,200)$	Models with convexity (BM25, TFIDF Std.) + Models with length normalization (BM25, TFIDF Norm.) + Traditional models(Boolean Chronological order, Boolean Reverse Chronological order)
Wilkie and Azzopardi (2013)	Discussions on the relationship between retrievability and the retrieval effectiveness, across different forms of document length normalization.	2.4 M	TREC test collections were used Associated Press (AP), Aquaint (AQ), and DotGov (DG)	692 K	Two-term most frequent words	$c=(5,10,20,50,70,100)$ $\beta=(0.5,1.0,1.5,2.0)$	Length normalization models (Okapi BM25 and DFRs PL2)
Wilkie and Azzopardi (2014)	Discussions on query set volume and exploring trends when using smaller or larger query sets. Finding the number of queries for obtaining a comparable estimate of Gini and calculating correlation between retrievability scores and various numbers of queries	NA	Two TREC test collections: Aquaint (AQ) and TREC123 (T123)	250 K for each collection	Two-term most frequent words which occurred at least 20 times	$\beta=[0-1]$ and $c=(10,100)$	BM25, PL2, DPH, TFIDF
Traub et al. (2016)	Assessing retrievability bias using a newspaper collection by various IR models	69 M	TREC collection and a digitized newspaper archive	1 K	Collecting real query logs in a period of 2 months from the DL by taking permission	$c=10,100,1000$	BM25, LM1000, TFIDF

TABLE 2 (Continued)

Author (year)	Focus	Document set (D)		Query set (Q)		Parameters	IR model (i.e., algorithm)
		Volume	Type	Volume	Method of generation		
Wilkie and Azzopardi (2016)	Analyzing the effect of a new topic centric approach on the estimation of retrievability bias	NA	TREC collection, AP-50 topics selected	600 queries per topic	Generating a topic centric set by initially identifying the pool of documents that were judged for a topic. Two-term most frequent words are then extracted from each pool.	$c = 100$	BM25
Wilkie and Azzopardi (2017)	Discussions on the bias of better-performing IR systems toward the relevant documents compared with the non-relevant documents.	1.8 M	TREC collections, AP, AQ,T45	1.5 M	First, generating the collection of single-term words and then issue the queries to each of the different configurations (collection, retrieval model, parameter setting).	$c = 100$	BM25, PL2, and LM with Dirichlet Smoothing (LMD)
Wilkie and Azzopardi (2018)	Examines the relationship between various IR models to fielding, retrieval performance, and retrieval bias by changing the weights of documents fields (i.e., title, author, body, source, etc.) and exploring the effect of missing fields to find the most robust field.	NA	TREC collections, AP, AQ,T45	NA	Similar to Bache (2011)	$c = 100$	BM25
Roy et al. (2022)	Assessing retrievability of publications and datasets using a real-life DL	830 K	Integrated search system named GESIS Search	8146 + 24,310	Collecting real query logs from the DL's search log	$c = (10,100)$	NA

retrievable, however, if only one document in the collection was retrievable, G_D is one (total inequality).

$$G_D = 1 - \frac{2}{N-1}\left(N - \frac{\sum_{d=1}^{N} d \cdot R(d)}{\sum_{d=1}^{N} R(d)}\right). \tag{3}$$

The Gini coefficient has been employed for quantifying the retrievability bias over a collection in several studies introduced in Table 2 (Wilkie & Azzopardi, 2013, 2014, 2018). In the studies, the authors examine the impact of several factors on the bias, including algorithms, the indexing process, the features of data set, and the parameter settings.

A Lorenz curve is a graphical representation of retrievability inequality. Documents are sorted based on their ascending retrievability score and the cumulative score is considered. As shown in Figure 1, the graph depicts the distribution of cumulative retrievability scores based on the percentiles of the data set shown on the vertical axis. The Gini coefficient can be calculated by the areas between the curve and the diagonal line (area A) and the area below the curve (area B) using = A/(A + B). Gini is zero when the curve lies on the equality line (area A is zero) and equals one when there is no area B on the plot.

3.2.2 | Other inequality metrics

The Hoover index is another metric for identifying the level of inequality formulated as Equation (4) for our retrievability scores ($R(d)$).

$$H_D = 0.5 \frac{\sum_{d=1}^{N} |R(d) - \overline{R}|}{\sum_{d=1}^{N} R(d)}. \tag{4}$$

The Atkinson index, formulated as Equation (5) is a normative inequality metric by applying an inequality-aversion coefficient (ε) to weight retrievability scores. We use a default value for this coefficient in this study ($\varepsilon = 0.5$). Guerrero (1987) details the formula and discusses the effects of the ε parameter on the Atkinson index.

$$A_D = 1 - \left(\frac{\sum_{d=1}^{N} \frac{R(d)}{\overline{R}} \cdot \left(1 - \frac{R(d)}{\overline{R}}\right)^{\varepsilon}}{N}\right)^{\frac{1}{\varepsilon}}. \tag{5}$$

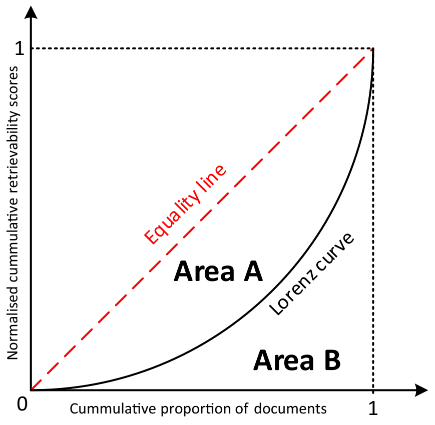


FIGURE 1 Lorenz curve and its relation to Gini coefficient.

Similar to Gini, the Hoover and Atkinson indices lie between zero (total equality) and one (total inequality). The Gini coefficient is a classic and widely used measure, while the Hoover index focuses on the redistribution. The Atkinson index offers flexibility through the aversion parameter, allowing to tailor the measurement to any specific research. We used the two indices to complement and check the Gini coefficient by offering different perspectives on inequality.

3.3 | Generating the query set

To compute retrievability, it is necessary to have a large number of queries and to estimate the probability of each query. The queries provided in most offline test collections Sanderson (2010) are not sufficient in number to allow an accurate estimate of retrievability to be calculated. Bache (2011) suggested that an estimated subset $Q' \in Q$ can be created artificially and for the probability of each query to be set at the same constant value: $Pr(q') = \frac{1}{|Q'|}$. This allows the rewriting of the retrievability function as Equation (6), where $\widehat{R}(d)$ denotes an estimate for document d.

$$\widehat{R}(d) = \sum_{q' \in Q'} f(\delta(q', d), \theta). \tag{6}$$

Among the studies listed in Table 2, some discuss how the design of the query set can improve retrievability estimates. Bashir and Rauber (2010) focus on the selection of relevant search queries for prior-art search in patent retrieval. They expand the query set using a pseudo-relevance feedback method. Following the study of Azzopardi and Vinay (2008b), Bache (2011) introduces an algorithm for selecting an adequately large query set, Q'. Instead of selecting single-term queries, they collect two-term words from the patent data set. By using Boolean operators (AND and OR), they examine the effect of these operators on the retrievability of the data set within various IR models. Here, the Gini coefficient and the mean frequency of retrievability for each corpus are calculated. Bache (2011) states that for comparing these metrics for different collections, the number of queries should be determined based on the number of documents in each data set by considering the "document-to-query ratio" as constant. Wilkie and Azzopardi (2014) find an efficient number of queries for obtaining a comparable estimate of Gini. They test several query sets with various numbers of queries to determine how correlated are the retrievability scores.

As can be seen in Table 2, all studies employ query sets that are sampled from document collections. However, the sets do not seem to be similar to the queries submitted by users.

3.4 | Summing up

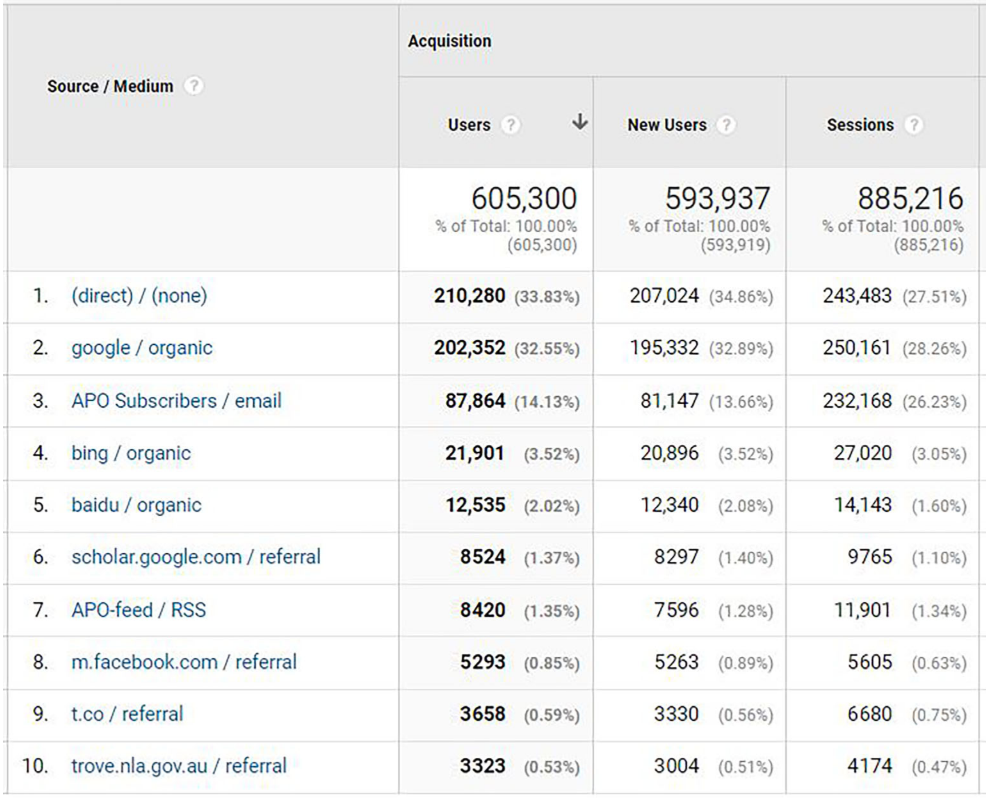
We see in the articles described in both the literature review and methodology section that there has been extensive examination of retrievability in a range of different data sets and contexts. However, most of the contexts have been in offline evaluation settings using pre-existing document collections. Also, the queries used in almost all of the past articles have employed the collection sampling technique first proposed by Azzopardi. In the next section, we detail our approach, which is distinct to past work.

4 | DATA

In this section, we describe the document collection, the set of queries, and the retrievability data used in our experiments.

The document collection consists of the summary pages collected from the "apo.org.au" DL. The library is the Analysis and Policy Observatory (APO), which is "a unique collection of material published by organizations (also known as grey literature) on any public policy issue—covering Australia, New Zealand and beyond."[2] At the time of this research, the APO contains over 36,000 documents, including articles, literature reports, in PDF and/or video formats. The data set receives 3 million+ page-views and 500,000+ downloads per year.[3] The documents in the collection are a mixture of content that is unique to APO as well as content that is replicated in other sources.

Figure 2 illustrates a typical report from the APO's GA platform in the same period of data collection for the channels used by users. The reports show that cumulatively, 39.5% of the landing pages were sourced from organic searches of four well-known web search engines:



Source / Medium	Acquisition		
	Users ↓	New Users	Sessions
	605,300 % of Total: 100.00% (605,300)	593,937 % of Total: 100.00% (593,919)	885,216 % of Total: 100.00% (885,216)
1. (direct) / (none)	**210,280** (33.83%)	207,024 (34.86%)	243,483 (27.51%)
2. google / organic	**202,352** (32.55%)	195,332 (32.89%)	250,161 (28.26%)
3. APO Subscribers / email	**87,864** (14.13%)	81,147 (13.66%)	232,168 (26.23%)
4. bing / organic	**21,901** (3.52%)	20,896 (3.52%)	27,020 (3.05%)
5. baidu / organic	**12,535** (2.02%)	12,340 (2.08%)	14,143 (1.60%)
6. scholar.google.com / referral	**8524** (1.37%)	8297 (1.40%)	9765 (1.10%)
7. APO-feed / RSS	**8420** (1.35%)	7596 (1.28%)	11,901 (1.34%)
8. m.facebook.com / referral	**5293** (0.85%)	5263 (0.89%)	5605 (0.63%)
9. t.co / referral	**3658** (0.59%)	3330 (0.56%)	6680 (0.75%)
10. trove.nla.gov.au / referral	**3323** (0.53%)	3004 (0.51%)	4174 (0.47%)

FIGURE 2 Source of landing pages during this study's selected date range (According to GA's definitions: User = "who have initiated at least one session during the date range." New Users = "first-time user." Session = "the period time a user is actively engaged with a website, app, etc. All usage data [Screen Views, Events, Ecommerce, etc.] is associated with a session").

Google, 32.55%; Bing, 3.52%; Baidu, 2.02%; Google Scholar, 1.37%. Access via search on the APO's website accounted for 33.83% of accesses. This pattern of accesses is a feature common to many digital libraries Ćirić and Ćirić (2021).

Each document has a catalogue summary page detailing title, authors, organization, subject area, and description. Each summary page is identified by a URL that finishes with "/node/" plus a unique document ID. The majority of the APO's website pages (almost 99%) are summary pages, the remainder of which are dedicated to detailing the website's and the APO. An API is employed to extract the summaries from the APO's Content Management System (CMS). We utilized the entire summaries for our tests as Google uses them for the retrieval. The APO was selected for our studies, as it is a DL of sufficient size to enable study of this topic, and, thanks to the kind contribution of the APO, we were given access to logs that would enable an analysis of Google searches.

The set of all Google queries that retrieve at least one APO document in the top 1000 Google results was obtained via the GSC API. The queries, covering 3 years, were extracted APO's Google Analytics tool between the dates November 17, 2018 and November 16, 2021. The resulting *GSC data set* contains 3.9 million records. Each record contained the query, the landed URL, and four other columns:

1. the impressions of APO documents (the number of times any APO document URL appeared in search results viewed by searchers, not including paid Google Ads search results),
2. the number of clicks on the APO document URLs,
3. CTR (=Clicks/Impressions × 100), and
4. the average rank position of the APO's document in the search results.

Table 3 details the descriptive statistics for the GSC data set columns in terms of queries related to APO documents. We use the average position of each URL, related to the document d and searched by query q', as $\delta(q',d)$ (see Equation (6)). For each query, we computed the number of words and characters in each query finding that on average, the queries for the documents are 3.3 words in length (20.5 characters). Figure 3 graphs query lengths in relation to CTRs, and confirms that although most of the queries contain two words (see Figure 3a), longer queries are associated with higher CTRs,[4] see Figure 3b. The past approach of generating sampled queries for retrievability experiments composed of only one or two words may not be ideal.

Examining the GSC data set we found that 27,729 unique documents in the APO DL were retrieved by at least one query. This number contrasts with the 36,329 documents held in the library. It would appear that 24% of the corpus (8600 documents) was not retrieved. An examination of page-views for each of the non-retrieved documents showed the documents were visible from the APO's CMS. We took a sample of 110 documents (over 1% from the 8600 non-retrieved) and searched on Google manually for the document title (with quotation marks) accompanied by the word "APO." We wanted to understand if the document was in Google's index. We found that 93 of the documents were retrieved, leaving 17 that were not (15% of the sample size), the documents were not returned by the search engine. To further examine the non-retrievability of these documents, we also checked several other related queries (by including the document's author name, organization, and a part of description) but were unable to retrieve the documents. To ensure that there was not some wider issue with the documents, we searched for all 17 manually on Bing, Duckduckgo, and Yahoo. All of the documents were findable on at least one of the other search engines. Therefore, of the 8600 non-retrieved documents, we assumed that 15% (1290 documents) are non-indexed documents on Google. Therefore, in any of the following comparisons with the results of the simulation studies, we ignore the non-indexed documents and only consider the remaining set (35,039 Google retrievable documents).

5 | RETRIEVABILITY ANALYSIS

In this section, we compare the results of retrievability scores gained by various parameters of the metrics with the simulation results reported in Azzopardi and Vinay (2008b) (see Section 5.1). According to our GSC data set, the Equation (6) is reformulated as Equation (7) to consider the impressions ($=Imp(q')$, i.e., the number of times document d is retrieved in the searches). Table 3, Column Impressions—Docs declares that for 50% of the queries the impression is equal to one or two if we only consider the Google retrievable documents.

TABLE 3 GSC data set statistics.

	Impressions	Clicks	CTR (%)	Average position	Character length	Query length
Mean	6.7	0.05	0.4	42.2	20.5	3.3

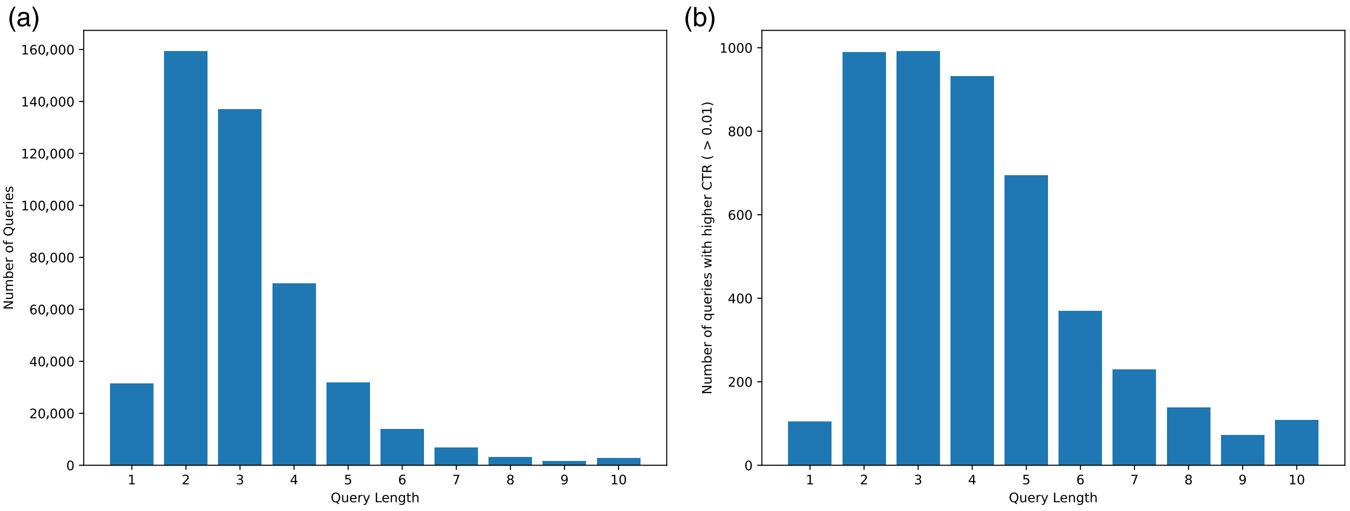


FIGURE 3 Query length, measured in words, in the GSC data set.

$$\widehat{R}(d)=\sum_{q'\in Q'}^{n} Imp(q')\times f(\delta(q',d),\theta). \qquad (7)$$

We compare the retrievability scores gained for various subsets (according to the attributes of the data set) and provide some applications for the calculated scores. We evaluate the consideration of Google real queries that includes various query lengths and compare the result with the result gained in the simulation studies considering the generated single- or two-word query sets. Finally, by assuming the CTRs gained from the GSC data set as query weights, we calculate the retrievability scores again and compare the results as per equally weighted queries with the corresponding results obtained by considering the new weights.

5.1 | Real versus simulated retrievability scores

For the first analysis, we calculate the retrievability score for each document ($\widehat{R}(d)$), using Equation (1) for the cumulative-based metric with $c=20,40,60,80$ and Equation (2) for the gravity-based metric with $\beta=0.5,1.0,1.5,2.0$, considering the set of queries Q'. We selected the parameters of the metrics according to the previous studies, aimed at comparing our results, gained by real data, with their simulation results.

Azzopardi and Vinay (2008b) examine the distribution of the retrievability scores in the shape of heat-maps, see Figure 1 in their article. For comparison, we also produce similar heat-maps with the same parameters of c and β, see Figure 4. Documents with zero or very low $\widehat{R}(d)$ values (less retrievable) are shown as white zones. In contrast, yellow or red zones indicate more retrievable documents with higher $\widehat{R}(d)$ values. It can be seen that increasing c means that the users investigate more pages on Google search results. Thus, for the cumulative-based metric, a larger portion of the document set is retrieved (see Figure 4a) and consequently, the white zones decrease. For the gravity-based metric (Figure 4b), increasing β results in poorer retrievability scores for some documents (as documents with more average positions are penalized more), and consequently less red zones are seen in the map. These trends correspond with the trends shown in the simulation study.

Examining the heat-maps depicted in fig. 1 of Azzopardi and Vinay (2008b), the white areas occupy ~20% of total in the cumulative-based and 40% in the gravity-based heat-maps. However, in our heat-maps, the white areas, shown in Figure 4a,b, are substantially larger. The authors of the previous article also note that in their simulation analysis, considering a rank cut-off of 100, over one-third (33%) of documents were not retrieved. Our number of documents with $\widehat{R}(d)\approx 0$ in $c=100$ show greater values (58%). This means that, by using real data, more zeros and low retrievability scores would be gained. It should be noted that in the previous studies, by applying artificial queries on all documents, it is more likely to obtain a rank for a document ($\delta(q,d)$) because the queries are selected from the documents.

For investigating the general trend of retrievability scores across all documents, following Azzopardi and Vinay (2008b), we calculate Pearson correlation coefficients between the retrievability values of cumulative- and gravity-based models, as shown in Table 4. Compared with the simulation result, we see similar trend (higher correlations with smaller c or β). Moreover, the trend of retrievability scores may change more when higher cutoff values are considered

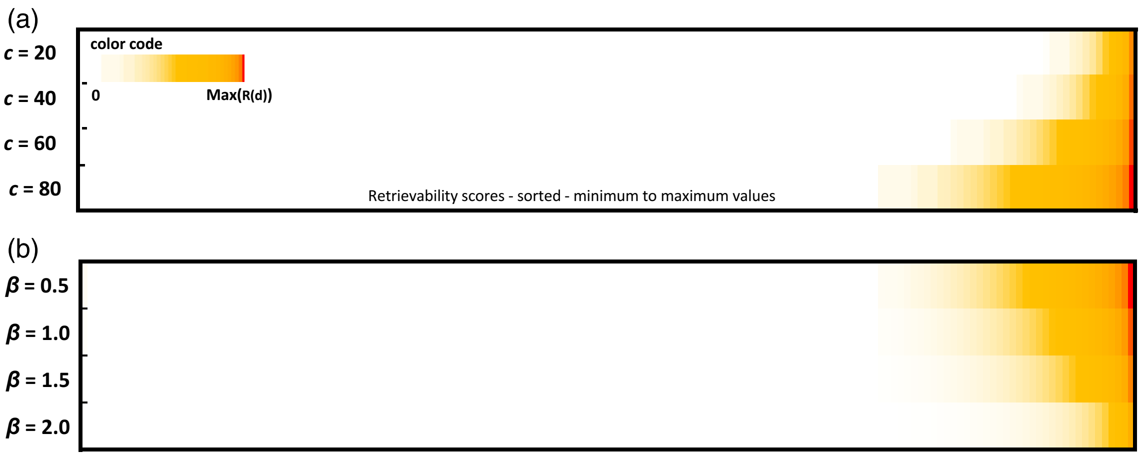


FIGURE 4 Heat-map of retrievability scores for the Google retrievable document set.

TABLE 4 Pearson correlation coefficient between retrievability scores estimated for different parameters of metrics.

		Cumulative-based metric			Gravity-based metric		
		$c = 40$	$c = 60$	$c = 80$	$\beta = 1.0$	$\beta = 1.5$	$\beta = 2.0$
$c = 20$	Real data	0.92	0.90	0.87			
	Simulation	[0.97,0.98]	[0.95,0.96]	[0.93,0.95]			
	Discrepancy	4.7%	4.9%	5.8%			
$\beta = 0.5$	Real data				0.89	0.80	0.73
	Simulation				[0.95,0.96]	[0.88,0.90]	[0.85,0.86]
	Discrepancy				5.4%	10.1%	13.4%

Note: For the simulation, since the values are computed for various IR models in Azzopardi and Vinay (2008b) (see Table 1), a range of values is shown. To calculate discrepancy, the difference between the simulation and real data is calculated based on the minimum correlation value gained in the simulation results.

(e.g. compare the discrepancy values for $c = 80$ with $c = 20$ and $c = 60$). In the gravity-based metric, we see more dissimilarity between the simulation and real data results (compare the discrepancy values for $\beta = 2.0$ and $\beta = 1.5$ with $\beta = 1.0$).

We also calculated the correlation between the retrievability scores gained by $c = 100$ with the results of $c = 80$ and $c = 200$ (0.999 and 0.941 were gained, respectively), which shows that the retrievability scores considering cutoff values more than 100 will gain identical results as per $c = 100$. So, we consider $c = 100$ as the highest notable cutoff for further tests consistent with Azzopardi and Vinay (2008b).

For the examination of retrievability bias, we created a Lorenz curve as explained in Section 3.2. Figure 5 compares the Lorenz curve, drawn by Azzopardi and Vinay (2008b) by a simulation approach, with our curve gained by real data employing cumulative-based metric with $c = 20$. Regardless of IR algorithm, the simulation curves depicted in the related studies (Azzopardi & Vinay, 2008b, fig. 2; Bashir & Rauber, 2010, fig. 7.1) are less skewed than the curve we find with our DL and GSC data set. The curve calculated by Google retrievable documents results higher values of Gini coefficient (=0.79). We found the "Exact Match" IR model in Bashir and Rauber (2010) similar to our curve. Our result for the gravity-based metric shows a lower Gini value (=0.64) than cumulative-based (=0.79). Figure 5b also confirms that less than half of the APO's documents are ranked outside of the top 20 results (having zero retrievability scores considering $c = 20$).

We compare the inequality metrics of a document set, introduced in Section 3.2, with respect to different parameters of retrievability measures. The comparison gives us an insight into how the willingness of APO users to look further down the search result will affect retrievability inequality within the document set. We draw the metrics in larger scales in Figure 6 to highlight the effect of c and β parameters on the retrievability bias. Figure 6a illustrates that increasing c (the willingness of users to explore more search results) leads to more equally distributed scores (less Gini, Hoover, and Atkinson measures). The inequality values for $c = 100$ and $c = 200$ confirms that there is no remarkable change in the

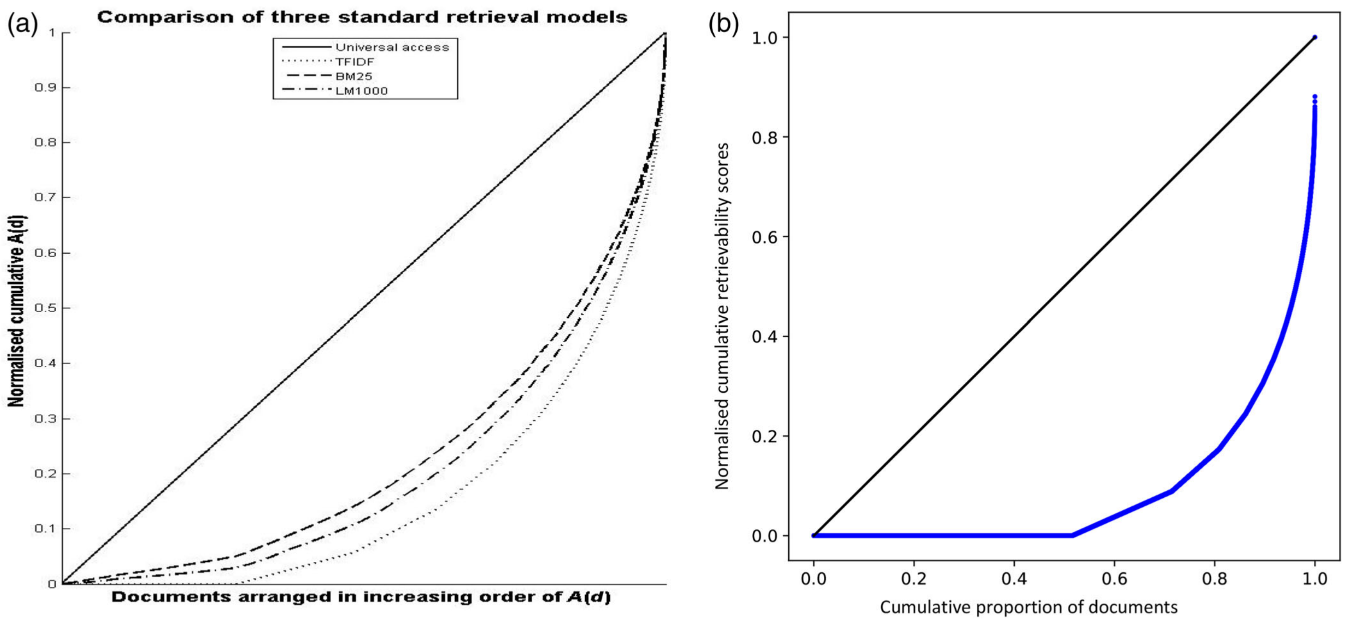


FIGURE 5 Lorenz curve of retrievability scores for the simulation and real data, using cumulative-based metric with $c = 20$.

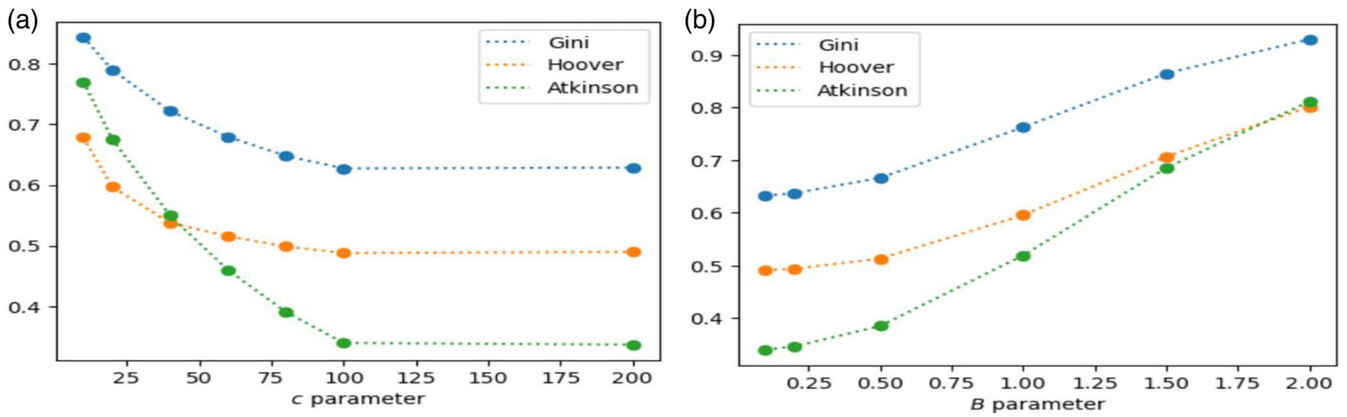


FIGURE 6 Inequality measures with respect to various parameters of retrievability for the real data.

distribution of retrievability considering the cutoff values greater than $c = 100$. Similar trends can be observed in the simulation results of Bashir and Rauber (2010) and Bache (2011). Figure 6b specifies a minimum $\beta = 0.1$ that results in the least retrievability bias. The comparison of the both figures indicates that the retrievability bias is more sensitive to changing the cutoff parameter in the cumulative-based than β in the gravity-based. This shows the importance of selecting an appropriate cutoff value if we use the cumulative-based metric.

5.2 | Comparing the simulation with reality in more detail

Given the substantial difference between the results derived from the DL logs and those published in past retrievability articles, we compared retrievability scores from a simulation and from the GSC data set in more detail. We created three sets of scores computed from:

Set 1: The queries of the logs of the DL and the retrievability scores (r_1, r_2) taken from the GSC. Here, the cutoff and beta parameter values were set at $c = 100$ and $\beta = 0.5$, respectively.

Set 2: We took an information retrieval system based on BM25 ranking Robertson and Zaragoza (2009). The system indexed the documents from the DL. The set of queries we used for the retrievability experiments were bi-gram queries (the classic retrievability simulation) extracted from documents of the DL collection. We measured retrievability at different cutoffs of c from 1 to 100.

TABLE 5 Pearson correlation coefficient between retrievability scores of documents.

		Set 1: Google, GA Qs			
		C	r_1	r_2	Gini
Set 2	BM25, bigram Qs	1	*−0.032*	*−0.025*	0.516
		5	*−0.028*	*−0.023*	0.480
		10	*−0.023*	*−0.017*	0.451
		20	*−0.024*	*−0.019*	0.405
		50	*−0.024*	*−0.020*	0.332
		100	*−0.028*	*−0.023*	0.273
Set 3	BM25, GA Qs	1	*0.287*	*0.292*	0.793
		5	*0.212*	*0.219*	0.633
		10	*0.180*	*0.187*	0.569
		20	*0.153*	*0.160*	0.511
		50	*0.128*	*0.134*	0.440
		100	*0.116*	*0.122*	0.389
		Gini	0.931	0.913	

Set 3: We used the same BM25 ranker indexing the same documents, but this time using the queries taken from the GSC data set.

We measured the Pearson correlation between retrievability scores of documents across Sets 2 and 1 (the top part of Table 5) and between Sets 3 and 1 (the bottom part of Table 5).

We found that the bigram query retrievability scores (Set 2) were not correlated at all with retrievability scores from the GSC data set (Set 1), while the GSC queries using the BM25 ranker (Set 3) were mildly correlated with the scores from Set 1; the BM25 ranker has some similarity in behavior as the Google ranker, but the different query set produce very different retrievability behaviors. Lower cutoff values of *c* were found to result in higher correlations. This makes sense because a search engine such as Google is only likely to retrieve one or two items from a given DL to present in its own ranking.

Turning to the Gini scores also reported in Table 5, we observe that Google is much more biased in what it retrieves from the collection than BM25. We see this in the high Gini scores showed at the bottom of the table (Google) compared with the Gini scores on the right side of the table (BM25). Comparing the Gini scores of Sets 2 and 3, we can see that the GSC queries produce more biased results than the bigram queries.

From these results, we conclude that the query simulation process as used in earlier retrievability articles does not correspond well with the reality of retrievability of a web search engine over the contents of a DL. The differences observed are due to both the ranking algorithm of the search engine and the queries issued by users. Next, we examine how attributes of documents in our DL might impact on how they are ranked.

5.3 | Document attributes

Using the retrievability scores from the DL, we subdivided the documents in the DL based on attributes. There have been many examinations of potential biases of search engine algorithms over the years, including questions of partisanship (Robertson et al., 2018) or rankers learning from biased clicks (Yue et al., 2010). For the content of this DL, we chose to examine temporal factors (Campos et al., 2014) and a traditional concern of search engines, document length (Singhal et al., 1996)—a topic Azzopardi & Vinay, 2008b also examined in the context of retrievability. Here, we consider overall document length.

As can be seen in Table 6, retrievability scores decline almost monotonically as the publication age of the documents increases. However, as can be seen, there is also a correlation between $\overline{L}_D$ and $\overline{R}_{100}$, the longer the documents are, the lower the retrievability. As document length also correlates with document age, it is not possible to separate the factors that might be affecting retrievability.

6 | LIMITATIONS OF STUDY

To the best of our knowledge this is the first time that the question of retrievability has been tested on the content of a digital library via the searches of an external web search engine. However, it is worth noting that there are limitations to this study.

- This is a study of just one DL, there may be qualities of this particular library and its content that may be impacting on the results. In future work, we plan to re-run this study on other DLs.
- This is a study that is conducted on a query log that has recorded past interactions with the DL. As such, most of the analysis in this article are studies of correlation rather than studies of causation. This is an inherent feature of this style of study, but it is one that must be remembered. It should also be remembered that this is a log drawn from a web search engine that is regularly updated and that searches a

great many other sites. Over the many years of study that this log covers, the search engine will have been altered in a multitude of ways multiple times. Other site will have held similar or even duplicate content to some of the content in our DL and that content will have regularly changed. All these alterations introduce noise into a correlation analysis such as ours. The presence of the noise does not invalidate this study, however, as it highlights the high levels of variability that DL owners face when trying to see how their content is accessed. We also find that the results of our examinations, such as that shown in Table 5 show that signal can be observed in our methodologies. It is important to show a study such as this given the importance of the external web search engines in accessing DL content.

- The library focuses on policy documents from two countries, Australia and New Zealand. While there are collections within in the library that have a broader focus, as with all almost all DLs, there is a focus to the content. It is not clear how this limitation might impact the results given that we are comparing different conditions and testing the generalizability of past retrievability results nevertheless this aspect of the library should be noted.
- The documents held by the APO are not necessarily unique to this DL. The presence of documents duplicated on other DLs may well cause a level of noise in the measurements of retrievability on the APO. However, we do not see this aspect as a bug of our analysis. The owner of a DL will be interested in knowing how retrievable their content is on an external web search engine and this analysis delivers this understanding. However, it is important to understand that the retrievability that is being measured is the retrievability of the content as it is stored on one particular DL.
- While the APO is a well-used library, it is a relatively small composed of around 36,000 searchable documents. Although this might impact the generalizability of the results, we feel that the scale of the collection is modest studying retrieve ability on 36,000 documents is still a study that provides a scale of a notable size. Retrievability is measured on the rank position of individual documents. Even with a collection of this modest size, there are still thousands of documents whose rank position is being tested which provides a level of scale that we feel is sufficient for a valid result to be published. Note that the data set we collected contained millions of records representing a substantial number of queries that were used to access the documents of the APO.

TABLE 6 Average retrievability scores with $c = 100$ ($\overline{R}_{100}$) and average length of documents' descriptions ($\overline{L}_D$) per year.

Year	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005
$\overline{R}_{100}$	0.19	0.18	0.17	0.11	0.07	0.06	0.06	0.06	0.04	0.04	0.06	0.04	0.04	0.04	0.02	0.03	0.03
$\overline{L}_D$	38.0	40.1	40.2	45.2	40.8	55.9	48.2	48.0	51.2	50.2	49.0	49.8	54.0	60.8	59.0	57.9	57.3

7 | CONCLUSIONS AND FUTURE DIRECTIONS

In this article, we estimated the retrievability of the content of a DL as accessed through queries to a well-known web search engine. We employed two measures of retrievability: cumulative- and gravity-based metrics and calculated the corpus retrievability metrics: the Gini coefficient and average retrievability scores. In contrast with almost all past studies, we calculated retrievability using a query set obtained from a log of queries via the GSC. Our work allowed us to answer the following research questions:

1. What is the retrievability of the documents in a DL to users of an external web search engine?

- Comparing the heat-maps of the retrievability scores in Figure 4a,b against past work (Azzopardi & Vinay, 2008b) leads us to conclude that the distribution of retrievability scores in this work is more skewed than in past work. The retrievability of documents in the DL was highly variable and substantially more variable than found in the retrievability experiments of past work.
- A great many documents in the DL were not retrieved over the time period studied. Examination of these documents found that for almost all of them, they were indexed by the external search engine, they just had not been ranked in any of the queries submitted to the engine. User queries on an external web search engine can match on so many more documents than can be stored on a single DL.

2. Can we predict retrievability scores of documents in the DL based on the simulation methods detailed in past work?

- Comparing the results of this study with past retrievability research, it would appear that the simulations of past work, conducted on test collections and other DLs are a poor predictor of retrievability in a DL.
- Examining the correlations in Table 5, we see that the reasons for the differences found are due to differences in the query sets used and in the ranking algorithms. The key reason appears to be the simulated queries used in past work being a poor proxy of the actual queries users submit.

3. Can we identify which features correlate with higher retrievability in the external web search engine?

- An examination of document attributes found that the publication date and length of documents appeared to be both correlated with retrievability.

What this work has illustrated is the importance of understanding the diverse and uneven variety of queries submitted to search engines. Past studies relied on a simulation of queries that was drawn evenly from a corpus of documents. Such an approach has not been found to provide a good proxy of the queries that users submit to an external search engine. In this work, we were fortunate to be given access to a large query log of a large DL, however, such access is relatively rare. Many researchers wish to conduct research on sets of queries that are realistic. Attempts to create such query sets are many (Abolghasemi et al., 2023; Alaofi et al., 2023; Bailey et al., 2016; Dang & Croft, 2010). However, the goal of creating a set of representative queries for a given document collection, a task required for the work in this article is still not solved. How such a query set could be generated will be the focus of future work.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council Linkage Infrastructure, Equipment and Facilities (LIEF) program to fund the Linked Semantic Platforms for Policy and Practice, LE180100094. Open access publishing facilitated by RMIT University, as part of the Wiley - RMIT University agreement via the Council of Australian University Librarians.

ORCID

Hamed Jahani https://orcid.org/0000-0002-7091-6060

ENDNOTES

[1] Research on the bias present in personalization asks how much does personalization bias the documents that a user retrieves (Liu et al., 2020). Does personalization effectively make it impossible for someone, subject to such personalization, to retrieve certain documents in a collection? Despite much discussion of this form of bias (and the potential for it to create so called filter bubbles), there is little evidence that such a bias exists in many prominent search engines Bruns (2019).

[2] https://apo.org.au/about

[3] https://apo.org.au/page/browse

[4] We considered CTR > 0.01 as higher CTRs according to a recent survey showing that the average CTR for a search is >1% for all benchmark industries (Hubspot blog, 2022).

REFERENCES

Abolghasemi, A., Verberne, S., Askari, A., & Azzopardi, L. (2023). Retrievability bias estimation using synthetically generated queries. In The first workshop on generative information retrieval (GenIR@SIGIR23).

Alaofi, M., Gallagher, L., Sanderson, M., Scholer, F., & Thomas, P. (2023). Can generative llms create query variants for test collections? An exploratory study. In Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval (pp. 1869–1873).

Azzopardi, L., & Bache, R. (2010). On the relationship between effectiveness and accessibility. In Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval (pp. 889–890).

Azzopardi, L., & Vinay, V. (2008a). Accessibility in information retrieval. In European conference on information retrieval, Springer (pp. 482–489).

Azzopardi, L., & Vinay, V. (2008b). Document accessibility: Evaluating the access afforded to a document by the retrieval system. In Workshop on novel methodologies for evaluation in information retrieval, Citeseer (pp. 52–60).

Azzopardi, L., & Vinay, V. (2008c). Retrievability: An evaluation measure for higher order information access tasks. In Proceedings of the 17th ACM conference on information and knowledge management (pp. 561–570).

Bache, R. (2011). Measuring and improving access to the corpus. In Current challenges in patent information retrieval. Springer (pp. 147–165).

Bailey, P., Moffat, A., Scholer, F., & Thomas, P. (2016). Uqv100: A test collection with query variability. In Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval (pp. 725–728).

Bashir, S., & Rauber, A. (2010). Improving retrievability of patents in prior-art search. In European conference on information retrieval, Springer (pp. 457–470).

Bruns, A. (2019). *Are filter bubbles real?* John Wiley & Sons.

Campos, R., Dias, G., Jorge, A. M., & Jatowt, A. (2014). Survey of temporal information retrieval and related applications. *ACM Computing Surveys*, *47*, 1–41.

Ćirić, J., & Ćirić, A. (2021). The impact of the covid-19 pandemic on digital library usage: A public library case study. *Journal of Web Librarianship*, *15*, 53–68.

Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, *19*, 30–41.

Dang, V., & Croft, B. W. (2010). Query reformulation using anchor text. In Proceedings of the third ACM international conference on web search and data mining (pp. 41–50).

De Rosa, C., Cantrell, J., Carlson, M., Gallagher, P., Hawk, J., & Sturtz, C. (2011). Perceptions of libraries, 2010: Context and community. A report to the OCLC membership. Technical report. OCLC Online Computer Library Center, Inc., Dublin, Ohio.

Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C. P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., & Sølvberg, I. (2007). Evaluation of digital libraries. *International Journal on Digital Libraries*, *8*, 21–38.

Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics*, *54*, 306–316.

Guerrero, V. M. (1987). A note on the estimation of atkinson's index of inequality. *Economics Letters*, *25*, 379–384.

Hansen, W. G. (1959). How accessibility shapes land use. *Journal of the American Institute of Planners*, *25*, 73–76.

Hofmann, K., Li, L., & Radlinski, F. (2016). Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, *10*, 1–117.

Hubspot blog. (2022). What's a good clickthrough rate? new benchmark data for google adwords. https://blog.hubspot.com/agency/google-adwords-benchmark-data?__hstc=17269478.2a5ced89530dc4f76850b6b90b6023d7.1600739076779.1600739076779.1600739076779.1&__hssc=17269478.1.1600739076779&__hsfp=1672237820

Kelly, E. J. (2014). Assessment of digitized library and archives materials: A literature review. *Journal of Web Librarianship*, *8*, 384–403.

Li, R., Li, J., Mitra, B., Diaz, F., & Biega, A. J. (2022). Exposing query identification for search transparency. In Proceedings of the ACM web conference (pp. 3662–3672).

Li, Y., & Liu, C. (2019). Information resource, interface, and tasks as user interaction components for digital library evaluation. *Information Processing & Management*, *56*, 704–720.

Liu, J., Liu, C., & Belkin, N. J. (2020). Personalization in text information retrieval: A survey. *Journal of the Association for Information Science and Technology*, *71*, 349–369.

Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, *2*, 1–22.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, *3*, 333–389.

Roy, D., Carevic, Z., & Mayr, P. (2022). Studying retrievability of publications and datasets in an integrated retrieval system. In Proceedings of the 22nd ACM/IEEE joint conference on digital libraries, association for computing machinery, New York, NY (p. 9).

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, *4*, 247–375.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval (pp. 21–29).

Traub, M. C., Samar, T., Van Ossenbruggen, J., He, J., de Vries, A., & Hardman, L. (2016). Querylog-based assessment of retrievability bias in a large newspaper corpus. In 2016 IEEE/ACM joint conference on digital libraries (JCDL) (pp. 7–16).

Wilkie, C., & Azzopardi, L. (2013). Relating retrievability, performance and length. In Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval (pp. 937–940).

Wilkie, C., & Azzopardi, L. (2014). Efficiently estimating retrievability bias. In European conference on information retrieval, Springer (pp. 720–726).

Wilkie, C., & Azzopardi, L. (2016). A topical approach to retrievability bias estimation. In Proceedings of the 2016 ACM international conference on the theory of information retrieval (pp. 119–122).

Wilkie, C., & Azzopardi, L. (2017). Algorithmic bias: Do good systems make relevant documents more retrievable? In Proceedings of the 2017 ACM on conference on information and knowledge management (pp. 2375–2378).

Wilkie, C., & Azzopardi, L. (2018). The impact of fielding on retrieval performance and bias. *Proceedings of the Association for Information Science and Technology*, *55*, 564–572.

Yue, Y., Patel, R., & Roehrig, H. (2010). Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In Proceedings of the 19th international conference on world wide web (pp. 1011–1018).

How to cite this article: Jahani, H., Azzopardi, L., & Sanderson, M. (2024). Measuring the retrievability of digital library content using analytics data. *Journal of the Association for Information Science and Technology*, 1–16. https://doi.org/10.1002/asi.24886