Scientometrics
https://doi.org/10.1007/s11192-024-05175-0

Open access publications drive few visits from *Google Search* results to institutional repositories

Enrique Orduña-Malea[1] · Cristina I. Font-Julián[1] · Jorge Serrano-Cobos[1]

Received: 12 June 2024 / Accepted: 27 September 2024

Abstract

Given the importance of *Google Search* in generating visits to institutional repositories (IR), a lack of visibility in search engine results pages can hinder the possibility of their publications being found, read, downloaded, and, eventually, cited. To address this, institutions need to evaluate the visibility of their repositories to determine what actions might be implemented to enhance them. However, measuring the search engine optimization (SEO) visibility of IRs requires a highly accurate, technically feasible method. This study constitutes the first attempt to design such a method, specifically applied here to measuring the IR visibility of Spain's national university system in *Google Search* based on a set of SEO-based metrics derived from the *Ubersuggest* SEO tool. A comprehensive dataset spanning three months and comprising 217,589 bibliographic records and 316,899 organic keywords is used as a baseline. Our findings show that many records deposited in these repositories are not ranked among the top positions in *Google Search* results, and that the most visible records are mainly academic works (theses and dissertations) written in Spanish in the Humanities and Social Sciences. However, most visits are generated by a small number of records. All in all, our results call into question the role played by IRs in attracting readers via *Google Search* to the institutions' scientific heritage and serve to underscore the prevailing emphasis within IRs on preservation as opposed to online dissemination. Potential improvements might be achieved using enhanced metadata schemes and normalized description practices, as well as by adopting other actionable insights that can strengthen the online visibility of IRs. This study increases understanding of the role played by web indicators in assessing the web-based impact of research outputs deposited in IRs, and should be of particular interest for a range of stakeholders, including open access and open science advocates, research agencies, library practitioners, repository developers, and website administrators.

Keywords Academic search engine optimization · Institutional repositories · Spain · Universities · Altmetrics · Open access

✉ Cristina I. Font-Julián
crifonju@upv.es

[1] Department of Audiovisual Communication, Documentation and History of Art, The iMetrics Lab, Universitat Politècnica de València, Valencia, Spain

Published online: 12 October 2024

Introduction

Institutional repositories (IRs) allow individual researchers, especially those without repositories in their own disciplines, to make their research openly accessible (Jones et al., 2006) and permit institutions to operate as stewards of all types of digital materials produced by staff and students (Lynch, 2003; Pinfield et al., 2014). As such, IRs preserve intellectual property by collecting, describing, and making available digital assets on behalf of the institutions that generate them (Crow, 2002).

Considering IRs a core component of the institutions' academic communication systems, pivotal in disseminating research results (Ruiz-Conde & Calderón-Martinez, 2014), any decrease in IR visibility might limit the number of visitors to an IR's website and hinder the findability of publications, reducing their chances of being read and, eventually, cited. Simply stated, the visibility of IRs in search engine results is of particular relevance.

While various methods can be exploited to drive traffic to websites (e.g., direct, referral, paid search, organic search, social media, email, and display, etc.), clicks from a search engine's organic results (i.e., non-paid) are critical. For example, according to *Similarweb*, 69% of visits to *DSpace@MIT* originate from organic results.[1]

Among available search engines, *Google Search* today predominates, having cornered 91.61% of the search engine market share worldwide as of December 2023, according to *StatCounter*.[2] Thanks to its simplicity and speed, *Google Search* plays a leading role in searching for and finding scholarly material (DeRosa, 2010; Gardner & Inger, 2021; Griffiths & Brophy, 2005; Haglund & Olsson, 2008; Markland, 2006; Niu & Hemminger, 2012).

Thus, the presence of the content of IRs in *Google Search* results is pivotal in ensuring that this IR-hosted content is visited. This presence can be measured using a range of web-based metrics and, in this way, researchers are able to obtain broad insights into the role of repositories in promoting the use of scholarly information, while institutions are informed as to whether they need to implement actions of improvement.

However, measurement requires a highly accurate, technically feasible method for collecting and processing metrics related not only to the IR's primary domain name (e.g., riunet.upv.es) but also to each URL, especially those linked to bibliographic records (e.g., riunet.upv.es/handle/10251/105556).

The present study attempts for the first time to design and apply a search engine optimization (SEO) method to ascertain the visibility of IRs in *Google Search* and, to do so, it employs Spain's national university system as a baseline.

The rest of the paper is structured as follows. "Research Background" section describes academic SEO (A-SEO) and highlights its applications in relation to institutional repositories. "Method" section defines the SEO-based metrics and describes the data collection and cleaning processes. "Results" section presents our main findings, including the number of URLs ranked for each IR, the number of search expressions that result in these URLs being ranked, and the number of visits that the ranked URLs drive to their respective IR websites. URLs linked to publications are described according to their year of publication, language, type, and subject. "Discussion" section outlines the practical implications of the results obtained, provides recommendations for repository managers to enhance IR visibility, and

[1] https://www.similarweb.com/es/website/dspace.mit.edu/#traffic-sources.

[2] https://gs.statcounter.com/search-engine-market-share.

discusses the limitations of the method. "Conclusions" section draws together the main conclusions to be reached by the study.

Research background

Online searches have changed the way people (including researchers) learn and read (van Dijck, 2010). At the core of this disruptive event lies the search engine (Enge et al., 2015).

Search engines display a limited number of results for a given query, ordered by relevance, the latter being determined by a unique algorithm specific to each search engine that exploits a range of different factors (Lewandowski, 2023). Thus, users tend to click on the results that appear on the first search engine results page (SERP) (Höchstötter & Lewandowski, 2009; Malaga, 2008; Pan et al., 2007), while results ranked outside the top 100 are unlikely to be seen and even less likely to be clicked on. This generates competition between websites in their efforts to attract attention.

The fact that a web page is ranked among the top 100 has two possible and co-existing explanations: natural—e.g., the content is relevant, newsworthy, etc.—and artificial—e.g., the content has been created or promoted to optimize its relevance in relation to certain search expressions (i.e., organic keywords). This has given rise to SEO, that is, the design and application of strategies oriented to driving web traffic to websites from search engines via organic search results for targeted search terms (own definition, inspired by Davis, 2006; Enge et al., 2015; Ledford, 2015; Serrano-Cobos, 2015).

The application of SEO practices to scholarly publications (Academic-SEO or A-SEO) can likewise be defined as the creation, publication, dissemination, and promotion of scholarly literature in a way that makes it easier for search engines to crawl and index it, favoring its appearance in the first positions of the SERP for the most significant number of search terms possible (own definition, inspired by Beel et al., 2010).

A number of seminal works (Beel & Gipp, 2010; Beel et al., 2010) have described A-SEO and report the outcome of various tests of the effects of SEO practices on *Google Scholar*, identifying both ethical—e.g., quality metadata and rich abstracts—and unethical—e.g., link spam, content spam and duplicate spam—strategies.

The A-SEO literature has evolved into various branches among which we find studies focused on how ranking algorithms operate (Martín-Martín et al., 2017; Rovira et al., 2019, 2021); on the SERPs generated by academic-related search terms (Gonzalez-Llinares et al., 2020); and on scholarly objects, including theses and dissertations (Coates, 2014), journals (González-Alonso & Pérez-González, 2015; Lopezosa & Vallez, 2023), research institutions (Park, 2018), and universities (Kaur et al., 2016; Olaleye et al., 2018).

The A-SEO scholarly literature has also turned its attention to repositories in an effort to determine how many records hosted by IRs are actually indexed on search engines. In general, they typically find low indexing ratios. For example, Arlitsch and O'Brien (2012) reported the low indexing ratios of US repositories in *Google Scholar*, while Orduña-Malea and Delgado López-Cózar (2015) found neither *Google Search* nor *Google Scholar* to be accurate and representative of the content available in Latin American IRs. Similar results were reported by Alhuay-Quispe et al. (2017) when analyzing Peruvian IRs, while Yang (2016) concluded that PDF files not supplemented with metadata were not crawled, and so not discovered, by search engines.

However, the above findings are based on tailored searches (i.e., by title or URL mention) and, as such, fail to show how visible the repositories are, regardless of the search term used. To address these shortcomings, the present study proposes and applies a method based on collecting and cleaning data from professional SEO tools equipped with functionalities and metrics that can exhaustively characterize the presence of IR websites on *Google Search* and approximate the web traffic that the search engine results actually drive to the IR websites.

Method

This study employed a three-step methodological approach: identifying the Spanish IRs, collecting web metrics (at the repository- and keyword-levels) using an SEO tool, and describing the bibliographic records collected. Each step is detailed below.

The first step involved locating all the Spanish IRs. To do so, the official websites of all of Spain's higher education institutions (HEIs) registered in the *Registry of Universities, Centers, and Degrees* (RUCT)[3] were manually accessed as of October 2022 to check for the existence of a repository. Additionally, the *ROAR* and *OpenDOAR* databases were consulted. The exercise yielded 86 HEIs with a total of 73 repositories (84.9% of the institutions), considering *CEU* (*Cardenal Herrera*, *San Pablo*, and *Abat Oliva*) and *Europea* (*Madrid*, *Valencia*, *Canary Islands*) as unique multi-site HEIs sharing the same IR.

The second step involved collecting web-based metrics for each repository using an SEO tool. We opted for the premium version of *Ubersuggest*[4] given its ease of use and additional features. Interestingly, for our purposes here, *Ubersuggest*, which has been previously used in the A-SEO literature (Dadkhah et al., 2022), offers the possibility of analyzing domain names and obtaining a wide range of metrics related to the search terms for which a domain name appears in the top 100 results (i.e., organic keywords). It also ranks specific URLs.

Ubersuggest was set up to obtain data from the Spanish search engine market.[5] (Note, however, that as SEO tools only operate with domains or subdomains, all the repositories with URL syntaxes, including subdirectories and dynamic URLs, were excluded from the analysis.[6])

Each repository's domain name was included as the domain under analysis (e.g., riunet.upv.es), Spanish was marked as the language, and Spain was marked as the country. Then, the "Keywords by Traffic" and "Top Pages by Traffic" features were selected to collect all the search terms for which a given repository's domain name appears in the top 100 positions and all the domain name's URLs with a ranking, respectively. Additionally, a pair of metrics (number of search terms and number of visits) were also collected at a global scale.

We also chose to design a new indicator—keyword strength—to measure the relevance of a search term for an IR, establishing a minimum monthly volume of searches (1000) and a minimum number of visits to the repository (10), and ranking the repository URL in a relevant position on the SERP (top 10). All these metrics were collected using *Ubersuggest*.

The third step involved describing each repository's URLs ranked in the top 100 positions for at least one search term in *Google Search*. This process involved cleaning and normalizing all the URLs, for which the Debug-Validate-Access-Find (DVAF) method was

[3] https://www.educacion.gob.es/ruct.

[4] https://neilpatel.com/es/ubersuggest.

[5] The selection of one country/language market is complimentary in SEO analyses.

[6] The following URL was discarded: biblioteca.nebrija.es/cgi-bin/repositorio.

followed (Orduña-Malea et al., 2023). The protocols (e.g., http, https) and all URL parameters (i.e., the characters to the right of the "?" symbol) were removed from each URL. All duplicate URLs and URLs pointing to different manifestations of the same record (e.g., full text, bibliographic description) were then merged.

This process yielded a range of URLs, some of them related to informative content (e.g., repository's homepage, search pages, statistics, group or researcher profiles, comments, reviews, tags, categories, recent submissions), others directly related to publications (bibliographic records and full-text publications), which include persistent identifiers (i.e., PID-based URLs). The analysis of PID-based URLs is relevant as the latter reflect the presence of links to publications in *Google Search* results, and help to determine whether the IRs are indexed because of the publications hosted or because of other content.

A *Python* script was designed to access automatically the PID-based URLs and to extract the DC metadata embedded in the corresponding HTML page. Among the available metadata, the year of publication (*DCTERMS.issued*), type of document (*DC.type*), language (*DC.language*), and subject (*DC.subject*) were statistically analyzed. Three repositories were discarded as they did not include DC metadata in their bibliographic records.[7]

The document types included in the *DC.type* metadata fields were categorized into 26 broad subjects due to the impossibility of determining the meaning behind some categories (e.g., "Conference objects" might contain presentations, posters, conference papers or conference abstracts), and the use of different terms in application to the same document type (e.g., Bachelor's thesis, *TFG*, or final degree project).

As for the *DC.subject* metadata field, the process found identical terms in lower/upper case (e.g., "Nursing Care" and "nursing care"), with/without diacritics (e.g., "Educación" and "Educacion"), with/without spaces (e.g., "Counting" and "Counting"), singular/plural (e.g., "Farmhouse" and "Farmhouses"), numerical codes at the beginning or end of the subject (e.g., "6201 Architecture"), all/part cases ("coastal zone management" and "coastal zone management—Mexico"), equivalent terms in different languages (e.g., "educación social", "educació social" and "social education"), and even descriptive phrases (e.g., "biological control of plant diseases").

We opted to conduct the thematic analysis by characterizing each publication based on its keywords instead of counting their frequency. *ChatGPT v4* was employed to preprocess the data for October 2022 as a sample (143,306 records with metadata) and classify them into broad thematic areas, allowing the inclusion of up to two broad disciplines in each publication to consider interdisciplinary publications. Two prompts were designed and tested (see supplementary material). The results obtained were subsequently reviewed manually.

The data were subsequently analyzed to extract metrics at the repository level. Specifically, the search terms making each repository visible in *Google Search* (i.e., number of organic keywords), the web traffic generated by those searches (i.e., number of visits), and the repository coverage (i.e., number of bibliographic records per repository) were computed. Table 1 offers detailed information on all the metrics collected, including the quantified event, the unit of measurement, the metric, the corresponding performance indicator, and the different filters considered.

The measures were applied monthly for three months (October, November, and December 2022), from which we obtained 217,589 unique bibliographic records and 316,899

[7] zaguan.unizar.es, addi.ehu.es, and summa.upsa.es.

Table 1 Repository-level metrics

Event	Unit of measurement	Metric	Indicator	Scope	Filters	Geo	Time span
Rank	Search terms	Number of search terms	Visibility	Search terms for which at least one repository URL appears in the top 100 search results, excluding ads	All keywords and all URLs	Global	12 months
					All keywords and PID-based URLs	Spain	3 months
					Strong Keywords and PID-based URLs	Spain	3 months
Rank	Bibliographic records	Number of records	Coverage	Records hosted in the repository that appear in the top 100 results	All records for all keywords	Spain	3 months
					Records with at least 10 visits	Spain	3 months
					Records with no visits	Spain	3 months
Visit	Clicks	Number of visits	Audience	Visits from *Google Search* results to the repository	Visits to all URLs	Global	12 months
					Visits to all URLs	Spain	3 months
					Visits to PID-based URLs	Spain	3 months

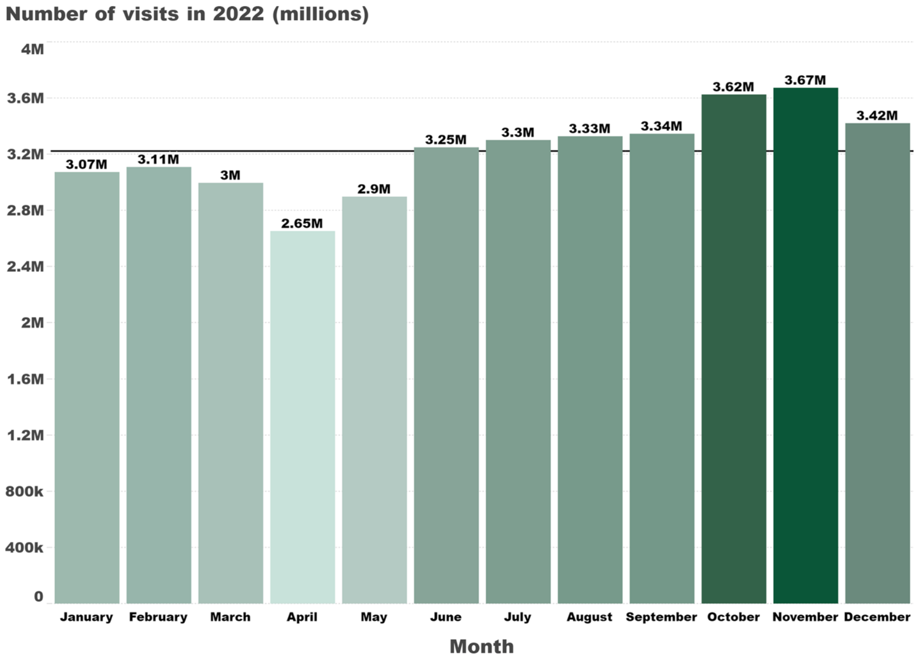

Fig. 1 Number of visits received by Spanish institutional repositories in 2022. Source: Based on data obtained from *Ubersuggest* and created with *Scimago Graphica* (https://www.graphica.app)

unique search terms. While the metrics related to the PID-based URLs are limited to these three months, the remaining repository-level metrics cover the whole of 2022.

Finally, the variability rates of the bibliographic records and search terms were designed to measure data variability. They were computed as the average of new records/keywords in month X concerning the records/keywords visible in month X-1.

Results

Repositories

In 2022, Spain's IRs received, on average, around 3.2 million monthly visits derived from clicks on the top 100 organic search results in *Google Search*, with October (3,624,923 visits) and November (3,672,485) figuring as the most active months and April (2,651,577) the least (Fig. 1). The total number of visits received from *Google Search* during the year amounted to 38,662,526.

While data corresponding to a broader span of years are needed to determine whether the results in Fig. 1 present a degree of seasonality, the fact that visibility dropped in April and was maintained in the boreal summer months (July and August) was, however,

Table 2 Global audience and visibility of Spain's institutional repositories in 2022

Repository domain name	Audience (global)		Visibility (global)		
	Visits (sum)	Visits (month Avg.)	Top 100 terms (month Avg.)	Top 10 terms (month Avg.)	Top 3 terms (month Avg.)
eprints.ucm.es	3,842,367	320,197	507,841	10,596	937
rua.ua.es	3,494,298	291,192	395,696	10,424	995
dadun.unav.edu	2,317,350	193,113	195,642	3574	250
riunet.upv.es*	2,012,325	167,694	299,033	6481	659
oa.upm.es*	1,582,906	131,909	269,646	7449	546
addi.ehu.es	1,516,526	126,377	149,962	1646	219
repositori.uji.es	1,507,663	125,639	105,437	2539	306
digitum.um.es	1,471,070	122,589	182,128	3828	260
e-spacio.uned.es	1,453,702	121,142	220,744	3990	296
upcommons.upc.edu*	1,444,263	120,355	309,747	5930	545
ddd.uab.cat	1,193,282	99,440	233,277	3620	341
gredos.usal.es	1,058,971	88,248	249,209	2959	211
idus.us.es	1,045,856	87,155	283,281	3045	213
repositorio.uam.es	1,024,808	85,401	264,510	2927	212
diposit.ub.edu	918,196	76,516	172,183	3730	335
uvadoc.uva.es	911,626	75,969	227,482	3014	223
ruc.udc.es	849,973	70,831	201,986	2673	140
accedacris.ulpgc.es	712,777	59,398	150,200	2316	164
digibug.ugr.es	685,303	57,109	134,419	1889	155
zaguan.unizar.es	673,216	56,101	214,969	2178	181

*Technical universities. Source: Based on data from *Ubersuggest*

unexpected. Sporadic falls in the number of visits to *Google Search* might, in part, explain these results.[8]

The number of visits to the IRs presents a skewed distribution, with 14 repositories (18.9%) obtaining more than one million visits each from *Google Search* in 2022. The IRs of the *Universidad Complutense de Madrid* (UCM) and the *Universitat d'Alacant* (UA) had the most accumulated visits in 2022 (Table 2).

The ranking achieved by UCM's repository can be attributed to the size of the university: the UCM is the third largest in Spain with 66,144 students enrolled in 2022/2023 (according to the *Integrated University Information System* [SIIU]) and the second largest in terms of scientific productivity with 107,217 publications (according to *Scopus*). Yet, the presence of medium-sized (UA, UPV) and small universities (UNAV) at the top of the ranking indicates that institutional size and productivity alone cannot account for the number of visits received from *Google Search*.

The UCM and UA also hold the repositories with the highest visibility in terms of organic keywords in *Google Search*'s SERPs (top 100, top 10, and top 3). Indeed, the Spearman correlation between the number of search terms for which a repository

[8] According to SEMRush (semrush.com), the number of organic visits to google.es dropped in March 2022 (from 47.6 million in February to 19.1 million in March), increasing again in April.

Table 3 Coverage, audience and visibility of Spain's institutional repositories Based on data from *Ubersuggest*

Repository domain name	Coverage (Spain)			Audience			Visibility (Spain)		
	PIDs (All)	PIDs (10-visits)	PIDs (0-visits)	Visits (Global)	Visits (Spain)	Visits (PIDs)	Terms (T100)	Terms (PIDs)	Terms (Strong)
idus.us.es	4146	347	655	102,854	22,538	18,073	308,964	4758	49
roderic.uv.es	4035	148	2393	42,255	9796	8642	102,719	4739	30
dadun.unav.edu	3749	139	1970	224,717	8407	8017	212,451	4695	23
uvadoc.uva.es	3735	521	429	132,725	34,988	31,065	351,545	4660	61
helvia.uco.es	3637	81	2500	22,498	7460	7259	68,529	4791	7
e-archivo.uc3m.es	3637	153	1951	36,255	8758	7928	121,177	4797	18
digibug.ugr.es	3604	202	1161	71,849	16,344	14,995	166,264	4711	39
gredos.usal.es	3601	326	552	93,659	20,442	18,120	265,244	4706	50
digibuo.uniovi.es	3578	84	2313	40,336	6709	6106	94,713	4741	15
riunet.upv.es	3534	476	442	195,307	32,168	28,406	340,012	4620	80
addi.ehu.es	3470	156	2034	55,764	10,366	9198	150,962	4756	25
upcommons.upc.edu	3455	349	464	154,552	23,838	20,540	367,507	4676	62
ebuah.uah.es	3389	156	1649	43,584	8562	7992	116,305	4744	33
minerva.usc.es	3374	139	1828	44,953	10,562	8883	93,425	4767	29
zaguan.unizar.es	3299	204	658	46,053	14,676	12,328	187,676	4654	30
diposit.ub.edu	3281	173	1825	72,156	10,957	10,230	168,210	4682	35
repositori.uji.es	3236	202	1146	194,249	27,585	16,549	131,709	4625	51
rabida.uhu.es	3226	43	2564	13,389	2817	2617	46,591	4745	8
oa.upm.es	3156	349	386	159,356	23,567	21,460	289,552	4663	86
repositori.upf.edu	3143	94	2073	17,494	5668	5388	68,546	4824	17

All values are averages of those recorded for the three months of October, November, and December 2022

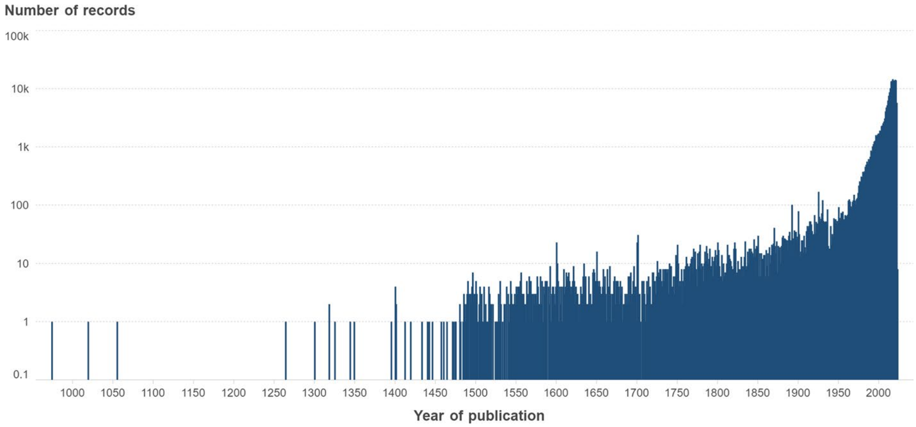

Fig. 2 Number of records by year of publication. Source: Based on the IRs' Dublin Core metadata and created with Scimago Graphica (https://www.graphica.app). Records taken from a sample of three months: October, November, and December 2023

appears in the top 100 results and the number of visits received after clicking on the link on the SERP is statistically significant ($R_s = 0.97\%$; p-value < 0.0001). However, some exceptions should be noted, most noticeably the technical universities, which have a relatively low number of total visits received from *Google Search* results compared to the number of search terms for which the repositories of these institutions are visible within the top 100 results.

The outcomes, however, vary significantly when the analysis is restricted to PID-based URLs (data from the Spain/Spanish filter). Thus, as Table 3 shows, two traditional, large, face-to-face institutions, the *Universidad de Sevilla* (US) and the *Universitat de València* (UV) achieve, on average, the highest coverage. Meanwhile, the IRs of the UCM and UA fall well down the rankings (26th and 30th, respectively).

Our analysis of the *Google Search* results from Spain yields a number of relevant findings. First, 84.9% of visits to Spanish IR websites come from PID-based URLs (i.e., publications are relevant in driving visits to the IRs). Second, visits from Spain in Spanish account for just 16.4% of the total visits from *Google Search* (i.e., there is a primary international audience).

The number of indexed publications that do not generate any visits to the repository is also quite remarkable (51.3%). Indeed, this proportion exceeds 80% for 26 IRs and is significantly higher in private universities. The UA (9%) and UCM (12%) are among the universities with the lowest percentage of bibliographic records with no visits, which may in part explain the results in Table 2.

These results corroborate the limited number of records with a minimum visit threshold (i10 visits index), shown in Table 3. Even though the correlation between the i10 visits index and the number of PID-based URLs is positive and strong ($R = 0.87$; p-value < 0.0001), this relationship is marked by a number of key exceptions. For example, the universities of *Córdoba* (UCO), *Oviedo* (UNIOVI), and *Huelva* (UHU) hold repositories with many PID-based URLs, but they have low i10 visit values (Table 3).

As for the vocabulary of the keywords that triggers the appearance of PID-based URLs, the results show an average of 3596 (SD = 1551) terms, with similar values being recorded by those IRs that enjoy a higher coverage (4600–4800 search terms). Otherwise, the number of "strong keywords" is limited (Mean = 20; SD = 24.6).

Records

Considering all 72 IRs analyzed herein, the number of PID-based URLs visible in *Google Search* results (top 100) for the Spain/Spanish filter totals 217,589. However, the actual number of publications is smaller, since a given publication might be deposited in more than one repository under a different "unique" identifier.

The distribution of publications according to the year of publication is shown in Fig. 2. This highlights the concentration of recent publications reflecting the increase in scientific production, the legal and institutional regulations that require the deposit of publications in Spain, and the annual deposit of students' academic output. Older publications illustrate the retrospective deposit of publications in IRs, even when these were not published by staff affiliated to the institution.

Most publications are written in Spanish (79.4%), followed by English (7.1%), and Catalan (4.1%). The presence of Spain's regional languages, including Catalan, Galician and Euskera, has grown significantly recently (Fig. 3). A plausible explanation for this is the increase in students' academic work and doctoral theses written in vernacular languages.

Journal articles constitute the most frequent type (29.7%), followed by students' academic work (27.8%) and doctoral theses (14.9%). While the distribution is similar to that of total records deposited in Spain's IRs according to *OpenDoar* data (Fernández, 2018), student academic work and theses are overrepresented (Fig. 4).

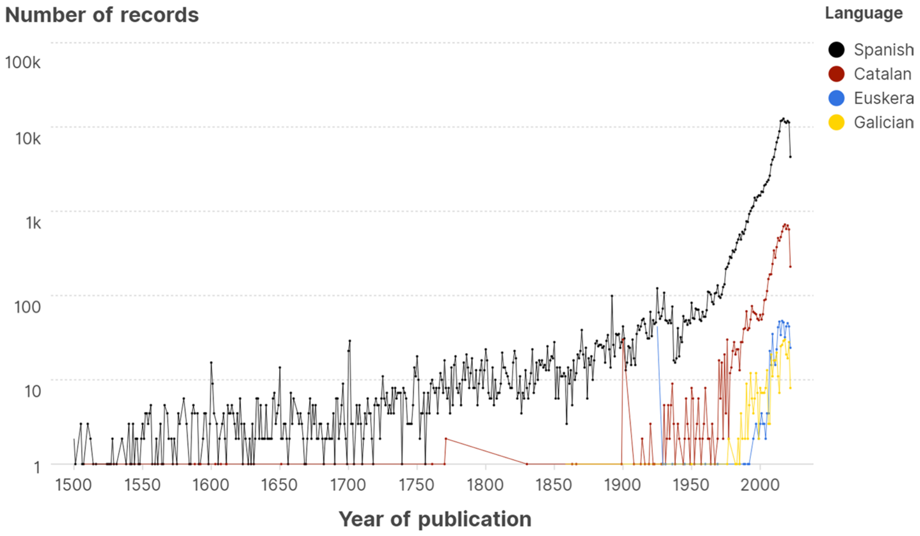

Fig. 3 Number of records per language. Source: Based on the IRs' Dublin Core metadata and created with Scimago Graphica (https://www.graphica.app). Records taken from a sample of three months: October, November, and December 2023

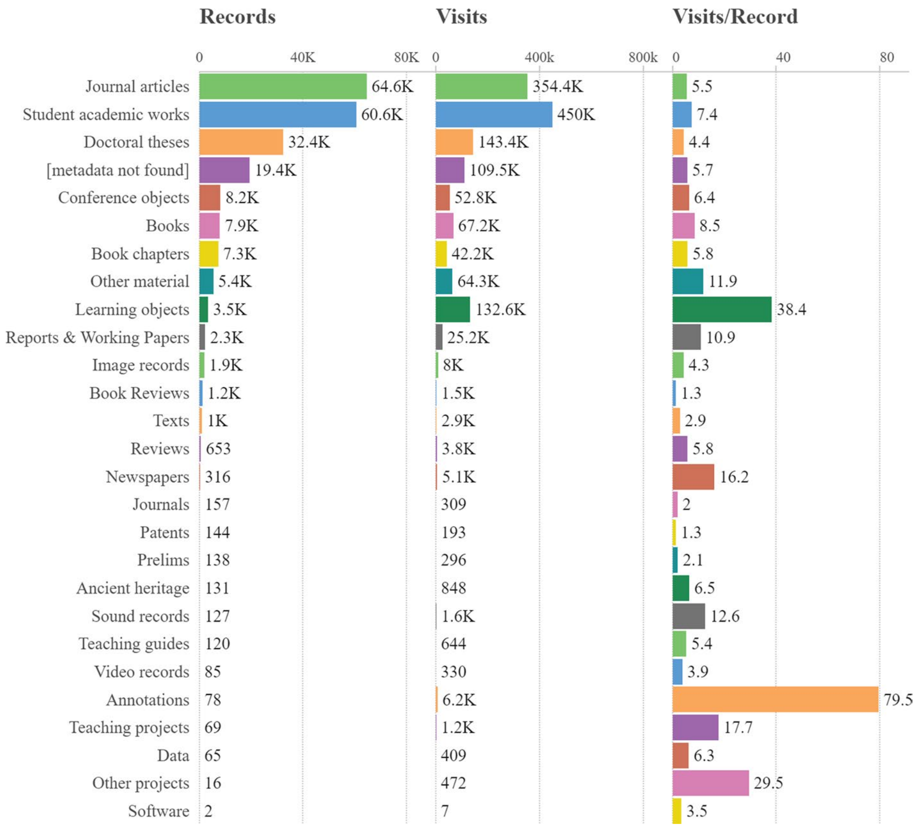

Fig. 4 Number of records and visits per document type. Source: Based on the IRs' Dublin Core metadata and created with Scimago Graphica (https://www.graphica.app). Records taken from a sample of three months: October, November, and December 2023. For each record, all visits from the three months are aggregated. Visits from Ubersuggest (Spain/Spanish filter used)

As for the accumulated number of visits by type, student academic output constitutes the most popular document type, with around 450,000 visits on average. Learning objects (38.4 visits per record on average) and Annotations (79.5 visits per page) also record a notable impact. Overall, our results point to the importance of the academic material and output created by students.

The audience attracted by students' academic work and by teaching materials is confirmed by the results reported in Table 4. This shows the 20 publications with the most visits accumulated during the three months of measurement, led by an undergraduate dissertation reporting the use of *Instagram* in the sports media (15,850 visits), followed by a conference paper (tagged as a journal article) on linguistics (13,798) and a language teaching text (also tagged as an article) for use by students wishing to perform a text commentary exercise (11,141). This discrepancy between metadata labels and document types is discussed in detail below.

No publications published in 2022 appear among the top twenty most visited records during the three months analyzed. This might indicate that the records require time to

Table 4 Most visited records from the Top 100 *Google Search* results

PID-based URL	Year	Type	Visits
uvadoc.uva.es/handle/10324/33101	2018	Student work	15,850
repositori.uji.es/xmlui/handle/10234/80348	1997	Journal article [*Conference paper*]*	13,798
digitum.um.es/digitum/handle/10201/28456	2012	Journal article [*Learning object*]*	11,141
rua.ua.es/dspace/handle/10045/61250	2017	Journal article	7310
upcommons.upc.edu/handle/2099.1/3319	2006	Student work	5506
riuma.uma.es/xmlui/handle/10630/20862	2021	Other material [*Learning object*]*	5326
eprints.ucm.es/id/eprint/45917	2018	Learning object	5086
eprints.ucm.es/id/eprint/45916	2018	Learning object	5023
dugi-doc.udg.edu/handle/10256/4195	2012	Conference object	4754
riuma.uma.es/xmlui/handle/10630/7299	2014	Other material [*Learning object*]*	4050
rua.ua.es/dspace/handle/10045/3834	2008	Report	3519
eprints.ucm.es/id/eprint/45915	2018	Learning object	3406
rua.ua.es/dspace/handle/10045/4298	2008	Learning object	3220
digibug.ugr.es/handle/10481/70294	2021	Book	3178
riunet.upv.es/handle/10251/30383	2013	Learning object	2990
eprints.ucm.es/id/eprint/45914	2018	Learning object	2936
repositorio.comillas.edu/xmlui/handle/11531/1038	2015	Student work	2896
repositorio.uam.es/handle/10486/681172	2017	Doctoral thesis	2834
minerva.usc.es/xmlui/handle/10347/7353	2011	Journal article	2715
digitum.um.es/digitum/handle/10201/86422	2020	Journal article	2478

*In brackets, the actual document type following manual inspection. Visits from *Ubersuggest* (Spain/Spanish filter used)

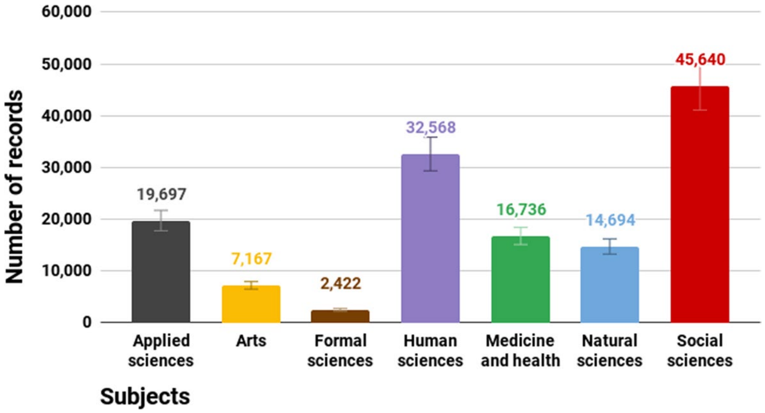

Fig. 5 Number of records per discipline. Source: Based on the IRs' Dublin Core metadata and created with Google Drive. Note: Records can be classified into more than one discipline

accumulate visits. However, data covering a broader time span are required to confirm this phenomenon.

Our analysis classified 112,091 records (73%) thematically. The disciplines with the most significant presence were the Social Sciences (40.7% of records) and the Human

Sciences (29%), while the presence of the Natural and Applied Sciences, as well as that of Medicine and Health, was smaller. The remaining disciplines can be deemed residual (Fig. 5).

Discussion

The present study has defined A-SEO metrics for non-aggregate (bibliographic records) and aggregate (repositories) online research objects and applied them to 217,589 bibliographic records and 316,899 organic keywords. As such, this is the first large-scale A-SEO analysis to be conducted of IRs to date and one that, moreover, provides comprehensive coverage of an entire national university system. In short, the study has served to enrich the evaluation of IRs using standard evaluation criteria (Serrano-Vicente et al., 2018).

Our findings indicate that many records deposited in the repositories are not ranked among the top positions in *Google Search* results and that most visits are generated by a small number of records.

Further studies are needed to determine whether these limitations are also evident in other countries. However, this effect may well be enhanced by the massive use of non-optimized repository software and the progressive transformation of *Google Search* into a centripetal website, offering the most significant amount of information possible to prevent the user from navigating to another site (Scolari, 2008).

Given the importance of *Google Search* in generating website traffic, these results point to the broad invisibility of scientific publications hosted by Spanish IRs to searches made in Spain, especially in Medicine, the Natural Sciences, and Engineering. This undermines considerably the dissemination function of the IRs and may limit the use and impact of the OA access hosted content and be a disservice to the academic reputation of Spain's universities.

For example, the *Universidad de Sevilla* is home to the repository with the largest number of PID-based URLs (4146, of which around 30% are journal articles), a number that pales in comparison with the 91,100 records that this institution had indexed in *Google Scholar* as of 3 January 2023 and the 120,300 total records deposited in the IR on that date,[9] and the 64,550 publications indexed in *Scopus*.

It should be borne in mind that the size of the repositories represents a fraction of a university's actual scholarly output due to low self-archiving rates, which is a consequence of various factors, including the use of thematic repositories and academic networking sites (Borrego, 2017; Xia, 2008), copyright conflicts and publishing practices that affect the percentage of OA repository deposits, despite the existence of institutional policies and national and supranational mandates (Abadal et al., 2013; De Filippo & Mañana-Rodríguez, 2022).

Low self-archiving rates combined with low indexation rates in *Google Search* means that the majority of publications from Spanish universities never appear in the top positions of *Google Search* results for searches conducted in Spain, at least as far as those versions hosted in the IRs are concerned. Indeed, this calls into question the role played by IRs in attracting readers to the institutions' scientific heritage via *Google Search*.

[9] https://guiasbus.us.es/ld.php?content_id=34832070.

A potential explanation for the low visibility recorded here is the general absence of repository website optimization, which would appear to include inadequate bibliographic descriptions and the creation of websites of limited usability and poor navigability.

The IRs' weak web authority (i.e., lack of external links from reputable sites)—an aspect widely recognized in the literature (Zuccala et al., 2008; Aguillo et al., 2010; Smith, 2012, 2013; Orduña-Malea, 2013; Fan, 2015; Orduña-Malea & Delgado López-Cózar, 2015; Aguillo, 2020)—can also lead to lower ranking positions in *Google Search* results, thus further reducing the chances of attracting visits.

Among the potential limitations of studies of this type, the use of inappropriate metadata schemes stands out, an issue first highlighted by Arlitsch and O'Brien (2012). Our study goes one step further by analyzing all PID-based URLs ranked in *Google Search* top results for all Spanish IRs. By accounting for the number of visits these results drive to the IRs' websites, we are able to corroborate not only the low indexation rates, but also the limited traffic generated to these repositories.

Arguably, the bibliographic records hosted in the IRs cannot compete in terms of ranking with other more popular web pages as regards the search terms used by users in *Google Search*, which we assume are likely to be more general and less specialized than search queries conducted in other bibliographic databases, such as *Google Scholar*. However, given the not insignificant number of visits that *Google Search* can drive to the IRs' websites, this issue should be addressed to make research results as visible to society as possible.

The best-positioned records are, to a large extent, academic studies and doctoral theses, mainly in the Social and Human Sciences. Given that the web pages of a given repository all use the same template, style sheet, and metadata scheme, it would appear that publications addressing social issues and the humanities, especially those written by students, employ a language that is less technical in nature, and which aligns better with the language employed by the majority of users of *Google Search*. This implies that more scientific works (or their metadata) should seek to include additional content that is more suited to a non-academic public, thereby favoring the transfer of their results to society.

Despite applications of SEO aimed at making academic content more readily available (Beel & Gipp, 2010), its use remains controversial. Indeed, some members of the academic community consider it a means to deceive academic search engines by artificially promoting low-quality content (Baeza-Yates, 2018), and, thus, it might undermine the relevance of search results.

SEO, it is claimed, takes advantage of the trust that users have in search engines (*European* Commission, 2016) and who would typically consider less credible works that are ranked lower (Ma, 2022) due to the relevance notion (i.e., ranking algorithm) built by search engines in their process of social platformization (Ma, 2023; van Dijck et al., 2018). This argument may well account for the few attempts made to date to improve or fund the online visibility of IRs.

Limitations

The method employed herein to collect bibliographic records and associated SEO metrics in *Google Search* presents a series of limitations, which we discuss below.

First, SEO analyses might be biased towards the larger universities. To check for a potential bias, we collected official staff data from SIIU and compared them with the SEO-metrics obtained (Table 5).

Table 5 Correlation matrix between Faculty staff, publications, and records indexed

Variables	1	2	3	4	5	6
1	1	0.88	0.76	0.60	0.68	0.71
2	0.88	1	0.81	0.74	0.75	0.75
3	0.76	0.81	1	0.75	0.93	0.89
4	0.60	0.74	0.75	1	0.72	0.80
5	0.68	0.75	0.93	0.72	1	0.93
6	0.71	0.75	0.89	0.80	0.93	1

All values are different from 0 with a significance level of alpha = 0.001

Legend: (1) Number of faculty staff (obtained from SIIU, course 2022/2023); (2) Number of publications (obtained from *Scopus*, historical data until 2022); (3) Number of records indexed in *Google Scholar* (obtained from *Transparency Ranking*); (4) Number of PID-based URLs ranked in *Google Search* (obtained from *Ubersuggest*); (5) Number of visits from *Google Search* (obtained from *Ubersuggest*); (6) Number of visits from *Google Search*, only from PID-based URLs (obtained from *Ubersuggest*)

Our results show that universities with the greatest number of research staff publish most and have the most records indexed in *Google Scholar* (see strong correlations of around 0.8). However, the largest universities do not present the highest number of PID-based URLs ($R = 0.6$). Likewise, although strong, the correlation between the number of publications and PID-based URLs points to a number of relevant exceptions ($R = 0.7$).

The presence of small universities with IR websites that enjoy greater visibility in *Google Search* than those of their larger counterparts might be attributed, among other potential factors, to better A-SEO strategies.

Second, the analysis of the repositories employs a URL-based approach. For this reason, some repositories cannot be analyzed since they do not fulfil the technical requisites (e.g., having a domain or subdomain) and, as such, they compromise data collection. The method is also subject to the problems associated with URL changes and redirections. For example, "eprints.ucm.es" was redirected to a different domain, that of "docta.ucm.es", several months after data collection. Furthermore, a multiplicity of PID-based URL syntaxes was detected due to a lack of standardization in the URL configuration process in *DSpace*, including URL aliases that hinder the process of identifying bibliographic records and collecting SEO metrics.

Third, while the method based on collecting DC metadata has its advantages (accessing all records *en masse* from the outside), it introduces a series of limitations. These are derived primarily from the total lack of standardization of these metadata in Spain's IRs, being implemented by default in *DSpace*, the most common repository software used by Spanish universities (Fernández, 2018).

The non-standard use of *DCTERMS.issue* (field dedicated to the publication date) and of *DC.date* (date of record incorporation into the repository) tags was detected, even within the same repository. Here, we used "DCTERMS.issue" metadata as the publication date when detected, while the DC.date metadata field was only used when the *DCTERMS.issue* metadata field was not present. As for document types, ambiguous categories (e.g., "Others"), synonymy (e.g., TFM, Master's thesis, final Master's work, coursework), and polysemy ("lectures" referring on occasions to conferences and, on others, to faculties' teaching presentations) were detected.

The analysis of subjects was even more challenging due to a complete absence of standardization, the field being used to insert keywords rather than subjects. *ChatGPT* solved the issue appropriately by characterizing each record with at least one subject. However, a manual inspection was still required, as the classification system failed to categorize some 63,711 records owing to the ambiguity of terms used (e.g., "leasing" was uncategorized rather than being associated with the "Social Sciences"). In other cases, the classification was erroneous due to misunderstandings (e.g., silviculture [forestry] was assigned to the Human Sciences) or an incapacity to categorize subjects written in languages other than those provided as examples in the prompt, written in Spanish. Although opting to use seven broad disciplines facilitated the classification process, the results should still only be considered a rough indication of subject.

Empty or incomplete records were also found: 22,746 (10.5%) records without a year of publication, 19,377 (8.9%) records without a typology, and 17,191 (7.9%) without a subject. Occasionally, this was due to the extraction parser, which could not correctly extract the metadata from the HTML. On other occasions, the content was void because the PID-based URLs corresponded to aggregates (journals, repository categories) without bibliographic metadata, or the repository was unavailable online. Finally, human mistakes need also to be taken into consideration, given the need for manual inspection to check validity.

Fourth, the use of web search engines for data collection also poses a series of technical challenges, especially as regards issues of coverage (Bar-Ilan, 2004; Vaughan & Thelwall, 2004; Lewandowski & Mayr, 2006; van den Bosch et al., 2016), the generation of unstable and rounded metrics (Font-Julián et al., 2018), and the bias of the user's search and location history (Badgett et al., 2015).

Fifth, the high degree of SEO data variability constitutes a notable limitation due to the high variability of web-based data between data samples. To verify this limitation, the

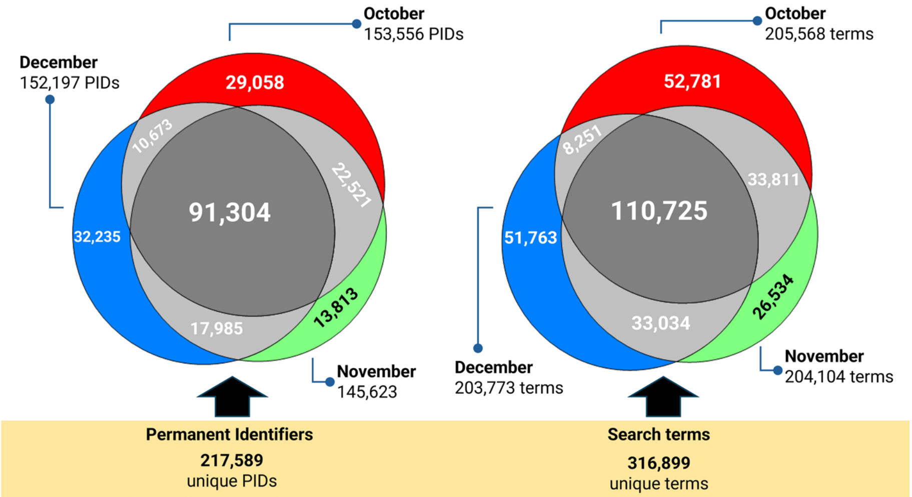

Fig. 6 Unique and shared records (left) and search terms (right) per month for Spain's institutional repositories. Source: Based on data from *Ubersuggest*

variability of PID-based URLs and keywords was measured between October and December 2022. This constitutes the first coverage analysis of its kind to be performed to date (Fig. 6).

Of the unique bibliographic records, 42% remained visible in the top 100 *Google Search* results across the three months in which evidence was collected, indicating a high degree of variability. For example, 32,235 (14.8%) records were identified uniquely in December, outside, that is, the top 100 results recorded in the previous two months (Fig. 6; left).

The search terms ensuring the records appeared in the top 100 *Google Search* results also presented a high degree of variability, with only 40% of the terms (110,725) appearing in all three months (Fig. 6; right).

For this reason, considering larger samples comprising more than 12 months' worth of data to verify possible annual patterns is recommended, as is obtaining more frequent samples (preferably on a weekly basis). However, the corresponding data collection and cleaning processes are technologically demanding.

Sixth, the data were obtained using a professional SEO tool (*Ubersuggest*). This means the methodology employed by the tool is dependent on the collection and computing of SEO data, which could differ from those offered by similar tools, such as *SEMRUsh, Sistrix,* or *Ahrefs*. However, this effect is minimized as all SEO tools collect data from the *Google API*, and tend to differ from each other in terms of other added-value features.

Seventh, the results reported herein are mainly limited to the Spain/Spanish area and *Google*'s search engine. However, other regions and languages should be included to obtain a broader view of the online visibility of IRs. Considering search engines beyond that of *Google Search*, especially for studies of IRs in other countries and demographic areas, is also necessary.

Conclusions

In light of our results, we conclude that Spain's institutional repositories need to improve their visibility in *Google Search* so as to facilitate the findability of publications. To achieve this goal, repository policies must be just as concerned with dissemination as they are with preservation. Indeed, actions are needed to boost the number of PID-based URLs that appear among search results, especially as far as journal articles in the Applied sciences, Engineering, and Medicine are concerned.

The method designed herein has proved to be somewhat time-consuming and is characterized by certain technical limitations in the collecting of metrics (dependent on the coverage of the tool) and the quality of the metadata (dependent on the policy adopted by each repository). Nevertheless, the method can be optimized, adapted, and extrapolated to analyze IRs from other regions, thematic repositories, publishers, journals, and any other online research resource and, as such, this constitutes a key contribution of the present study. Meanwhile, the consideration of other SEO tools, the design and application of additional SEO-based metrics, and the acquisition of standardized bibliographic descriptions constitute future challenges.

This work contributes to integrating SEO-based metrics into the array of alternative metrics, which should serve as valuable tools for assessing new dimensions of the dissemination and impact of online research objects, both individual (publications) and aggregate (repositories), which are of interest to the field of the Science of Science, in general, and to A-SEO, in particular.

To conclude, it is our belief that IRs should transition from infrastructures focused on hosting content to become true platforms of dissemination. A trigger here might be provided by the development of web-based and SEO-based indicators that move beyond the characteristics afforded by standard evaluation criteria. The present study, we believe, contributes to that end by facilitating the transition towards a new scenario in the quantitative and qualitative evaluation of IRs.

Acknowledgements The authors thank Daniela de Filippo and Ángel Delgado-Vázquez, who provided valuable comments on earlier versions of this manuscript, and Iain Robinson for proofreading the English text.

Author contributions Enrique Orduña-Malea: conceptualization, funding acquisition, methodology, writing-original draft. Jorge Serrano-Cobos: funding acquisition, writing-review & editing. Cristina I. Font-Julián: project administration, writing-review & editing.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Research funded by Grant GV/2021/141 (Regional Government of Valencia, Spain) and Grant PID2022-142569NA-I00, funded by MCIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe".

Data availability Available at https://doi.org/10.4995/Dataset/10251/209311. Additional data and information available at https://www.universeo.tech.

Declarations

Conflict of interest The authors have no competing interests to declare relevant to this article's content.

References

Abadal, E., Ollé Castellà, C., García, M. F. A., & Melero, R. M. (2013). Políticas de acceso abierto a la ciencia en las universidades españolas. *Revista Española De Documentación Científica, 36*(2), e007. https://doi.org/10.3989/redc.2013.2.933

Aguillo, I. F. (2020). Altmetrics of the Open Access Institutional Repositories: A webometrics approach. *Scientometrics, 123*(3), 1181–1192. https://doi.org/10.1007/s11192-020-03424-6

Aguillo, I. F., Ortega, J. L., Fernández, M., & Utrilla, A. M. (2010). Indicators for a webometric ranking of open access repositories. *Scientometrics, 82*(3), 477–486. https://doi.org/10.1007/s11192-010-0183-y

Alhuay-Quispe, J., Quispe-Riveros, D., Bautista-Ynofuente, L., & Pacheco-Mendoza, J. (2017). Metadata quality and academic visibility associated with document type coverage in institutional repositories of Peruvian universities. *Journal of Web Librarianship, 11*(3–4), 241–254. https://doi.org/10.1080/19322909.2017.1382427

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech, 30*(1), 60–81. https://doi.org/10.1108/07378831211213210

Badgett, R. G., Dylla, D. P., Megison, S. D., & Harmon, E. G. (2015). An experimental search strategy retrieves more precise results than PubMed and Google for questions about medical interventions. *PeerJ, 3*, e913. https://doi.org/10.7717/peerj.913

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM, 61*(6), 54–61. https://doi.org/10.1145/3209581

Bar-Ilan, J. (2004). The use of web search engines in information science research. *Annual Review of Information Science and Technology, 38*(1), 231–288. https://doi.org/10.1002/aris.1440380106

Beel, J., & Gipp, B. (2010). Academic search engine spam and Google Scholar's resilience against it. *The Journal of Electronic Publishing*. https://doi.org/10.3998/3336451.0013.305

Beel, J., Gipp, B., & Wilde, E. (2010). Academic search engine optimization (ASEO) optimizing scholarly literature for Google Scholar & Co. *Journal of Scholarly Publishing, 41*(2), 176–190. https://doi.org/10.3138/jsp.41.2.176

Borrego, Á. (2017). Institutional repositories versus ResearchGate: The depositing habits of Spanish researchers. *Learned Publishing, 30*(3), 185–192. https://doi.org/10.1002/leap.1099

Coates, M. (2014). Search engine queries used to locate electronic theses and dissertations. *Library Hi Tech, 32*(4), 667–686. https://doi.org/10.1108/lht-02-2014-0022

Crow, R. (2002). The case for institutional repositories: a SPARC position paper. *Washington, D.C.: Scholarly Publishing & Academic Resources Coalition*. Retrieved 8 June, 2024, from https://sparcopen.org/wp-content/uploads/2016/01/instrepo.pdf

Dadkhah, M., Rahimnia, F., & Memon, A. R. (2022). How frequent is the use of misleading metrics? A case study of business journals. *The Serials Librarian, 83*(2), 197–204. https://doi.org/10.1080/0361526x.2022.2145414

Davis, H. (2006). *Search engine optimization*. O'Reilly Media.

De Filippo, D., & Mañana-Rodriguez, J. (2022). The practical implementation of open access policies and mandates in Spanish public universities. *Scientometrics, 127*(12), 7147–7167. https://doi.org/10.1007/s11192-021-04261-x

DeRosa, C. (2010), *Perceptions of Libraries, 2010: Context and Community*. OCLC Online Computer Library Center, Dublin, OH. Retrieved 8 June, 2024 from https://files.eric.ed.gov/fulltext/ED532601.pdf

Enge, E., Spencer, S., & Stricchiola, J. (2015). *The art of SEO: Mastering search engine optimization*. O'Reilly Media, Inc.

European Commission. (2016). Online platforms: Report. *European Commission*. https://doi.org/10.2759/937517

Fan, W. (2015). Contribution of the institutional repositories of the Chinese Academy of Sciences to the webometric indicators of their home institutions. *Scientometrics, 105*(3), 1889–1909. https://doi.org/10.1007/s11192-015-1758-4

Fernández, T. F. (2018). Los repositorios institucionales: evolución y situación actual en España. In J. A. Merlo Vega (Ed.), *Ecosistemas del Acceso Abierto* (pp. 39–84). Ediciones Universidad de Salamanca. Retrieved 8 June, 2024, from https://gredos.usal.es/handle/10366/138583

Font-Julian, C. I., Ontalba-Ruipérez, J. A., & Orduña-Malea, E. (2018). Hit count estimate variability for website-specific queries in search engines. *Aslib Journal of Information Management, 70*(2), 192–213. https://doi.org/10.1108/ajim-10-2017-0226

Gardner, T., & Inger, S. (2021). *How readers discover content in scholarly publications: trends in reader behaviour from 2005 to 2021*. Renew Consultants. Retrieved 8 June, 2024, from https://renewconsultants.com/wp-content/uploads/2021/07/How-Readers-Discover-Content-2021.pdf

González-Alonso, J.., & Pérez-González, Y. (2015). Presencia en Google Scholar y en la WEB de la Revista Cubana de Plantas Medicinales. *Revista Cubana de Plantas Medicinales*, *20*(1), 1–13. Retrieved 8 June, 2024, from https://www.medigraphic.com/pdfs/revcubplamed/cpm-2015/cpm151a.pdf

Gonzalez-Llinares, J., Font-Julian, C. I., & Orduña-Malea, E. (2020). Universidades en Google: Hacia un modelo de análisis multinivel del posicionamiento web académico. *Revista Española De Documentación Científica, 43*(2), e260. https://doi.org/10.3989/redc.2020.2.1691

Griffiths, J. R., & Brophy, P. (2005). Student searching behavior and the web: Use of academic resources and Google. *Library Trends, 53*(4), 539–554.

Haglund, L., & Olsson, P. (2008). The Impact on university libraries of Changes in information Behavior among academic researchers: A multiple case study. *The Journal of Academic Librarianship, 34*(1), 52–59. https://doi.org/10.1016/j.acalib.2007.11.010

Höchstötter, N., & Lewandowski, D. (2009). What users see—Structures in search engine results pages. *Information Sciences, 179*(12), 1796–1812. https://doi.org/10.1016/j.ins.2009.01.028

Jones, R. E., Andrew, T., & MacColl, J. (2006). *The institutional repository*. Elsevier.

Kaur, S., Kaur, K., & Kaur, P. (2016). An empirical performance evaluation of universities website. *International Journal of Computer Applications, 146*(15), 10–16. https://doi.org/10.5120/ijca2016910922

Ledford, J. L. (2015). *Search engine optimization bible*. Wiley.

Lewandowski, D. (2023). Understanding search engines. *Springer*. https://doi.org/10.1007/978-3-031-22789-9

Lewandowski, D., & Mayr, P. (2006). Exploring the academic invisible web. *Library Hi Tech, 24*(4), 529–539. https://doi.org/10.1108/07378830610715392

Lopezosa, C., & Vállez, M. (2023). Audiencias amplias y visibilidad web: Posicionamiento de revistas académicas de comunicación en Google. *Index.comunicación, 13*(1), 153–171.

Lynch, C. A. (2003). Institutional Repositories: Essential infrastructure for scholarship in the digital age. *Portal: Libraries and the Academy, 3*(2), 327–336. https://doi.org/10.1353/pla.2003.0039

Ma, L. (2022). Metrics and epistemic injustice. *Journal of Documentation, 78*(7), 392–404. https://doi.org/10.1108/jd-12-2021-0240

Ma, L. (2023). Information, platformized. *Journal of the Association for Information Science and Technology, 74*(2), 273–282. https://doi.org/10.1002/asi.24713

Malaga, R. A. (2008). Worst practices in search engine optimization. *Communications of the ACM, 51*(12), 147–150. https://doi.org/10.1145/1409360.1409388

Markland, M. (2006). Institutional repositories in the UK: What can the Google user find there? *Journal of Librarianship and Information Science, 38*(4), 221–228. https://doi.org/10.1177/0961000606070587

Martín-Martín, A., Orduna-Malea, E., Harzing, A., & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics, 11*(1), 152–163. https://doi.org/10.1016/j.joi.2016.11.008

Niu, X., & Hemminger, B. M. (2012). A study of factors that affect the information-seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology, 63*(2), 336–353. https://doi.org/10.1002/asi.21669

Olaleye, S. A., Sanusi, I. T., Ukpabi, D. C., & Okunoye, A. (2018). Evaluation of Nigeria Universities websites quality: A comparative analysis. *Library Philosophy and Practice, 1717*, 1–14.

Orduña-Malea, E. (2013). Aggregation of the web performance of internal university units as a method of quantitative analysis of a university system: The case of Spain. *Journal of the American Society for Information Science and Technology, 64*(10), 2100–2114. https://doi.org/10.1002/asi.22912

Orduña-Malea, E., Alonso-Arroyo, A., Ontalba-Ruipérez, J. A., & Catalá-López, F. (2023). Evaluating the online impact of reporting guidelines for randomised trial reports and protocols: A cross-sectional web-based data analysis of CONSORT and SPIRIT initiatives. *Scientometrics, 128*(1), 407–440. https://doi.org/10.1007/s11192-022-04542-z

Orduña-Malea, E., & Delgado López-Cózar, E. (2015). The dark side of open access in Google and Google Scholar: The case of Latin-American repositories. *Scientometrics, 102*(1), 829–846. https://doi.org/10.1007/s11192-014-1369-5

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Park, M. (2018). SEO for an open access scholarly information system to improve user experience. *Information Discovery and Delivery, 46*(2), 77–82. https://doi.org/10.1108/idd-08-2017-0060

Pinfield, S., Salter, J., Bath, P. A., Hubbard, B., Millington, P., Anders, J. H., & Hussain, A. (2014). Open-access repositories worldwide, 2005–2012: Past growth, current characteristics, and future possibilities. *Journal of the Association for Information Science and Technology, 65*(12), 2404–2421. https://doi.org/10.1002/asi.23131

Rovira, C., Codina, L., Guerrero-Solé, F., & Lopezosa, C. (2019). Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft Academic. *WOS and Scopus. Future Internet, 11*(9), 202. https://doi.org/10.3390/fi11090202

Rovira, C., Codina, L., & Lopezosa, C. (2021). Language bias in the Google Scholar ranking algorithm. *Future Internet, 13*(2), 31. https://doi.org/10.3390/fi13020031

Ruiz-Conde, E., & Calderón-Martínez, A. (2014). University institutional repositories: Competitive environment and their role as communication media of scientific knowledge. *Scientometrics, 98*(2), 1283–1299. https://doi.org/10.1007/s11192-013-1159-5

Scolari, C. (2008). Online brands: Branding, possible worlds, and interactive grammars. *Semiotica, 2008*(169), 169–188. https://doi.org/10.1515/SEM.2008.030

Serrano-Cobos, J. (2015). *SEO: Introducción a la disciplina del posicionamiento en buscadores*. UOC Publishing.

Serrano-Vicente, R., Melero, R., & Abadal, E. (2018). Evaluation of Spanish institutional repositories based on criteria related to technology, procedures, content, marketing and personnel. *Data Technologies and Applications, 52*(3), 384–404. https://doi.org/10.1108/dta-10-2017-0074

Smith, A. G. (2012). Webometric evaluation of institutional repositories. *Proceedings of the 8th International Conference on Webometrics Informetrics and Scientometrics & 13th Collnet Meeting* (pp. 722–729). Seoul (Korea). Retrieved 8 June, 2024, from https://ir.wgtn.ac.nz/handle/123456789/18727

Smith, A. G. (2013). Web Based Impact Measures for Institutional Repositories. *Proceedings of the ISSI 2013 conference* (pp. 1806–1816). Viena (Austria). Retrieved 8 June, 2024, from https://ir.wgtn.ac.nz/handle/123456789/18790

Van den Bosch, A., Bogers, T., & De Kunder, M. (2016). Estimating search engine index size variability: A 9-year longitudinal study. *Scientometrics, 107*(2), 839–856. https://doi.org/10.1007/s11192-016-1863-z

Van Dijck, J. (2010). Search engines and the production of academic knowledge. *International Journal of Cultural Studies, 13*(6), 574–592. https://doi.org/10.1177/1367877910376582

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management, 40*(4), 693–707. https://doi.org/10.1016/s0306-4573(03)00063-3

Xia, J. (2008). A comparison of subject and institutional repositories in self-archiving practices. *The Journal of Academic Librarianship, 34*(6), 489–495. https://doi.org/10.1016/j.acalib.2008.09.016

Yang, L. (2016). Making search engines notice: An exploratory study on discoverability of DSpace metadata and PDF files. *Journal of Web Librarianship, 10*(3), 147–160. https://doi.org/10.1080/19322909.2016.1172539

Zuccala, A., Oppenheim, C., & Dhiensa, R. (2008). Managing and evaluating digital repositories. *Information Research: An International Electronic Journal, 13*(1), 3.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.