Factors for enhancing visibility in digital repositories: Metadata quality, interoperability standards, persistent identifiers, and SEO-GEO optimization

Danilo Reyes-Lillo
Universitat Pompeu Fabra, Spain
https://orcid.org/0000-0002-0141-8324
Cristòfol Rovira
Universitat Pompeu Fabra, Spain
https://orcid.org/0000-0002-6463-3216
Alejandro Morales-Vargas
Universidad de Chile, Chile
https://orcid.org/0000-0002-5681-8683

Reyes-Lillo, D., Rovira, C., & Morales-Vargas, A. (2025). Factors for enhancing visibility in digital repositories: Metadata quality, interoperability standards, persistent identifiers, and SEO-GEO optimization. In J. Guallar, M. Vállez, & A. Ventura-Cisquella (Coords). *Digital communication. Trends and good practices* (pp. 119-133). Ediciones Profesionales de la Información. https://doi.org/10.3145/cuvicom.09.eng

Abstract

In the context of open science and scholarly communication, enhancing the visibility of digital repositories is essential for maximizing the reach, impact, and discoverability of the content they host. This chapter explores five key strategies to improve repository visibility: optimizing metadata quality, enabling interoperability protocols, adopting persistent identifiers (PIDs), implementing search engine optimization (SEO) strategies, and embracing generative engine optimization (GEO). Through detailed analysis and practical recommendations, this work highlights how standardized metadata, controlled vocabularies, and persistent identifiers, such as DOI, Handle, and ARK, contribute to enhancing visibility. This chapter also emphasizes the importance of aligning repositories with evolving web technologies and AI-driven engines to ensure content remains accessible, traceable, and integrated into users' search experience.

Keywords

Digital repositories; Visibility; Metadata quality; Interoperability; Persistent identifiers; Search Engine Optimization; Generative Engine Optimization.

1. Introduction

In the era of open access and open science, digital repositories play a crucial role in disseminating knowledge. Enhancing their visibility not only broadens the reach of deposited content but also strengthens institutional impact and contributes to the democratization of information access. Various studies have shown that open-access publications, particularly those hosted in institutional repositories, tend to receive more citations and are more accessible than those restricted by paywalls (Piwowar et al., 2018; Swan, 2010).

Moreover, increased visibility in academic search engines, such as Google Scholar, and databases like OpenAIRE and CORE, enables repositories to comply with open-access mandates set by national and international funders (UNESCO, 2021). Strategies such as the proper implementation of standardized metadata, the use of persistent identifiers (such as DOIs and ORCID) and interoperability with other systems through protocols like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) are key to achieving greater exposure (OpenAIRE, 2020).

Therefore, actively working to optimize repository visibility is not merely a technical concern (Reyes-Lillo et al., 2025) but an institutional strategy to ensure that intellectual output fulfills its ultimate purpose: to be discovered, used, and cited by academic communities and society at large.

The following section analyzes five optimization techniques that can be employed to improve the visibility of digital repository content:

1. Metadata Quality Optimization.
2. Enabling Interoperability Protocols.
3. Adoption of Persistent Identifiers.
4. SEO Optimization of the Repository.
5. Generative Engine Optimization in Repositories.

2. Metadata quality optimization

Metadata standardization is crucial for ensuring the interoperability, visibility, access, and reuse of deposited content. A metadata strategy must consider not only technical aspects but also organizational and policy-related dimensions of the repository.

While the metadata model is often closely tied to the platform on which the repository is built, the schema must be adapted to local needs without compromising compatibility with international standards.

Table 1 presents a comparison of several digital repository software platforms and their underlying metadata schemas:

Table 1
Digital repository software and their base metadata schemas.

Software	Base Metadata Schema	Additional information
DSpace	Qualified Dublin Core https://github.com/DSpace/DSpace/blob/main/dspace/config/registries/dublin-core-types.xml	Uses a custom profile of Qualified Dublin Core; supports extensions like METS/MODS. From DSpace 7 onward, multiple schemas are supported.
Fedora	No predefined schema https://wiki.lyrasis.org/display/FEDORA6x/Data+Modeling	Employs RDF/Linked Data models; schema depends on implementation (MODS, DC, PREMIS, etc.).
EPrints	Configurable and extensible metadata schema https://wiki.eprints.org/w/Metadata#Metadata_Field_Types	Schema based on fields defined by the repository administrator. Can interoperate with other schemas via import/export.
InvenioRDM	JSON structure aligned with DataCite https://inveniordm.docs.cern.ch/reference/metadata	Schema conforms to DataCite's Metadata Schema v4.x with minor additions and modifications.
TIND	MARC21 https://www.tind.io	Based on MARC21 due to its foundation in INVENIO (developed by CERN). Offers various modules and can adapt to other formats.
Digital Commons	No specified schema; mappable to Dublin Core https://digitalcommons.elsevier.com/en_US/organization-content-planning/metadata-options-in-digital-commons	Declares a flexible metadata schema.
DataVerse	Uses various standard-compliant metadata schemas https://guides.dataverse.org/en/latest/user/appendix.html	Ensures interoperability and preservation through schemas like DDI, DataCite, Dublin Core, ISA-Tab, and VOResource, enabling structured export.

To initiate a metadata optimization strategy, it is recommended to begin with an initial assessment that includes at least the following elements:

- Audit of existing metadata: the goal is to review a representative sample of records to identify inconsistencies, formatting errors, empty or misused fields.
- Identification of metadata schemas in use: it is essential to identify both the base metadata schema and the various mappings and export capabilities to other schemas.

– Review of vocabularies and authority files: evaluate and validate the implementation of controlled vocabularies, such as thesauri, authority files, identifiers, and controlled lists, for item types, language fields, and date formats.

Once the assessment is complete, four fundamental processes are recommended to optimize repository metadata, which are described in the following section.

2.1. Data cleaning and refinement

This stage is essential for optimizing the visibility of content stored in a repository. High-quality metadata "should allow digital users to intuitively conduct the tasks such as identifying, describing, managing and searching data" (Ma et al., 2009, p. 1).

In this regard, data cleaning is a process that detects and corrects errors, inconsistencies, and incomplete fields, aiming to improve interoperability and user experience (Van-Hooland & Verborgh, 2015; Westbrook et al., 2012). This enhances accuracy and improves information retrieval through a system's search tools.

Among the various tools available to improve metadata quality in repositories, OpenRefine (https://openrefine.org) stands out. It is a powerful open-source tool for cleaning, transforming, and reconciling messy tabular data. OpenRefine is particularly useful for standardizing and enriching metadata in libraries, archives, and research datasets.

To use OpenRefine in the data cleaning process, it is essential to: 1) export metadata in a format compatible with OpenRefine, such as CSV, TSV, Excel, or JSON; and 2) ensure that each row represents a resource (such as a document or image) and each column corresponds to an element of the metadata schema. For example, using Dublin Core (DC), the columns might include *dc:title*, *dc:creator*, *dc:subject*, among others.

This approach enables batch cleaning of records with inconsistencies, typos, or formatting errors. Below is an example figure of DC records that could potentially be improved using tools like OpenRefine.

Figure 1

Example of DC records with potential for optimization through data cleaning.

dc:title	dc:creator	dc:subject	dc:date	dc:language
History of Science	Laura Gonzalez	Science; history	2020-06-10	eng
history of science	Laura González	science; history	10/06/2020	en
Artificial Intelligence	María Gómez	Technology, AI; Machine Learning	2023	english
Big Data		Massive data; analytics	2021-03-14	en
Machine Learning	Carlos Ruiz	machine learning, artificial intelligence	2024-11-25	en
Cybersecurity	Ana Lopez	information security	2021/05/03	English

It is worth noting that this process can be carried out regardless of the base metadata schema used by the system. To systematize metadata into a format accepted by OpenRefine, auxiliary tools such as MarcEdit (https://marcedit.reeset.net) can be used if the base schema is MARC21.

Once the records are imported, OpenRefine can correct inconsistencies such as unnecessary spaces, inconsistent capitalization, duplicates, and incorrect formats (especially in dates) as well as normalize terms using controlled vocabularies. Cells can be split, similar values

grouped, and transformations applied using GREL expressions[1]. Additionally, OpenRefine allows for the detection of missing values, format validation, and data reconciliation with external sources. Finally, the cleaned data can be exported in the desired format, ready for reuse or reintegration into the repository.

2.2. Implementation of controlled vocabularies

To enhance metadata with the goal of optimizing visibility and interoperability with other systems, controlled vocabularies play a significant role (Chipangila et al., 2024). On the one hand, they enhance the search experience within the repository itself by standardizing vocabulary in key fields, such as authors, language, and keywords. On the other hand, they enhance integration with external systems by using identifiers, controlled vocabularies, or lists, and predefined formats in key fields for interoperability.

To incorporate controlled vocabularies into digital repositories, the first step is to identify the metadata fields that could benefit from normalization. For example: authors, dates, language, keywords or subjects, and resource type.

Subsequently, it is necessary to select appropriate vocabularies to control specific fields. Below are some options categorized accordingly:

- To control subjects or keywords: Some alternatives include the Library of Congress Subject Headings (LCSH), the UNESCO Thesaurus, or the United Nations Bibliographic Information System (UNBIS) Thesaurus. Once the vocabulary to be integrated is selected, the next step is to integrate the thesaurus in a format such as SKOS (Simple Knowledge Organization System) or RDF (Resource Description Framework) into the platform. This integration will depend strictly on the platform and must ensure that subject or descriptor fields are mapped to the thesaurus terms (for example, the dc.subject field in Dublin Core or field 650 in MARC21). Finally, it is essential to verify that terms autocomplete correctly and are linked to their respective URLs.
- To control authorities, controlled vocabularies such as VIAF (Virtual International Authority File) or persistent identifiers like ORCID or ROR (Research Organization Registry) can be utilized. The integration process for these elements also depends strictly on the platform. Likewise, it must be ensured that the metadata schema fields related to authorities are correctly mapped to the vocabulary or list being integrated.
- To control item type: Various vocabularies exist to standardize item types and enhance interoperability, for example, with reference management systems. Notable examples include the DCMI Type Vocabulary (Dublin Core), COAR Resource Type Vocabulary, MODS Resource Types, Schema.org / CreativeWork Types, and OpenAIRE Guidelines Types, among others. After selecting the vocabulary, existing local values must be mapped to the chosen vocabulary, data entry forms should be adjusted to use controlled lists, and consistency of types across records must be ensured. Additionally, metadata export (e.g., in Dublin Core or XML) should be adapted to comply with interoperability standards and facilitate harvesting by aggregators. The process may include data validation, user interface adjustments, and testing to ensure that values are accurately reflected in both the internal administration interface (backend) and the public view of the repository.

1 GREL is a programming language designed to facilitate the organization, transformation, and querying of data in OpenRefine.

– To control key fields: normalizing key fields, such as language or publication date, is essential for standardizing metadata. Proper normalization improves the search experience within the repository and ensures effective data exchange with external services. For these fields, it is recommended to adopt ISO 8601 for date normalization and ISO 639 for language codes. This involves using the YYYY-MM-DD format for dates and two- or three-letter codes for languages. Implementation is achieved through controlled lists, automatic validation in data entry forms, and adjustments to the system's metadata templates, such as in Dublin Core, MARC21, or MODS (Metadata Object Description Schema). This ensures consistency, interoperability, and compatibility with external harvesters.

2.3. Periodic audits and adjustments according to regulatory changes

To carry out periodic metadata quality audits, it is recommended to establish a regular routine, such as quarterly or semiannually, in which representative samples of records are reviewed. During this review, consistency, completeness, correct use of controlled vocabularies, standardized formats (such as ISO 8601 and ISO 639-1) and the absence of typographical errors or duplicates should be verified. The use of automatic validation tools and quality report generation helps efficiently detect issues.

In response to changes in regulations or standards, it is essential to actively monitor updates in schemas such as Dublin Core, COAR, or OpenAIRE. When a modification occurs, the metadata mapping in the repository should be reviewed, and adjustments made to forms, vocabularies, or export templates. Staff should also be trained on the changes. Documenting each adjustment ensures traceability and facilitates future audits.

3. Enabling interoperability protocols

Enabling interoperability protocols in a digital repository is essential for increasing its visibility and reach. These protocols allow external services, such as aggregators, academic search engines, and national or regional portals, to automatically harvest the repository's metadata and content without manual intervention (Eells et al., 2024).

Thanks to this interoperability, the repository's records can be integrated into platforms such as OpenAIRE, BASE (Bielefeld Academic Search Engine), CORE (COnnecting REpositories), WorldCat, or even national or regional repositories, enhancing their discoverability for researchers, students, and the public. Additionally, it facilitates integration with other institutional systems and ensures that the content complies with open standards and open access policies.

There are several ways to promote interoperability through the following protocols:

– OAI-PMH Protocol: This protocol, based on HTTP and XML, enables other systems to harvest metadata from a repository automatically. It is widely used by aggregators such as OpenAIRE or BASE to collect information about available resources.
– REST API: This is an interface that allows other systems to query, create, or modify resources in a repository using HTTP requests, such as GET, POST, PUT, or DELETE. It is highly flexible and commonly used for integrations with external systems or custom applications.

– SWORD: This is a protocol that enables the remote deposit of content, such as articles, datasets, or theses, into a repository. It facilitates integration between publishing platforms, institutional systems, and digital repositories.

The activation of various protocols for interoperating with other systems strictly depends on the software and version being used. The available tools in each repository should be utilized to enable the API or SWORD, allowing for uploading or updating from external clients.

Likewise, each repository allows for the configuration of its OAI-PMH protocol, where the base URL of the service is defined and metadata is exposed in an appropriate format. In this context, metadata is essential and must be exposed correctly in formats readable by the protocol, such as Dublin Core, MARC, or DataCite.

The following table explains how to enable the OAI-PMH protocol according to each software platform:

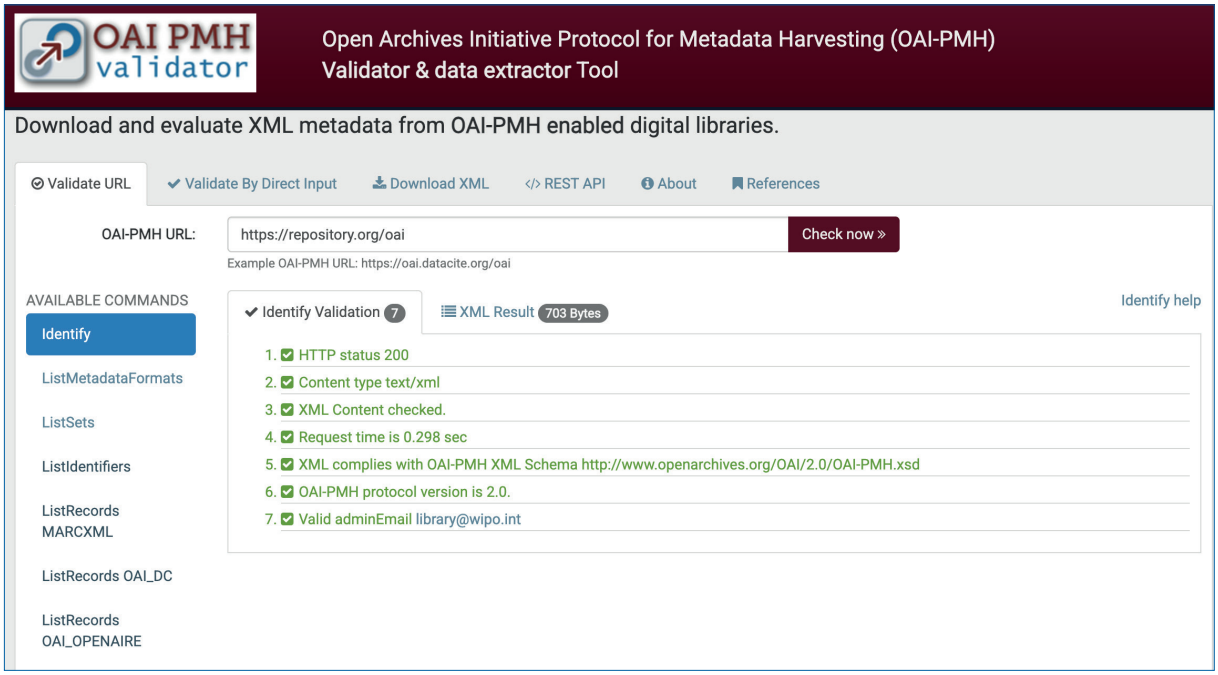
Table 2
OAI Protocol activation by software.

Software	How to Enable OAI-PMH?
DSpace https://wiki.lyrasis.org/display/DSDOC8x/OAI	Although OAI-PMH is enabled by default, it is necessary to verify that "oai.enabled=true" and "oai.path=oai" are set in local.cfg or dspace.cfg.
EPrints https://wiki.eprints.org/w/OAI	OAI support is enabled by default. You need to configure the base URL (oai.base_url), archive_id, sets, and any XSL stylesheets in cfg/cfg.d. Pay special attention to the oai.pl file.
Fedora https://github.com/saw-leipzig/foaipmh	Fedora 6 does not include native OAI-PMH support. It requires implementing an external endpoint (e.g., Django + foaipmh) connected to its REST API.
InvenioRDM https://inveniordm.docs.cern.ch/reference/oai_pmh/	OAI-PMH is enabled by default at /oai2d. From the admin interface, you can define sets and formats (oai_dc, oai_datacite).
TIND	From the admin panel, go to 'OAI Repository Admin', enable the provider, and define sets.
Digital Commons https://digitalcommons.elsevier.com/integration-preservation/digital-commons-and-oai-pmh	OAI support is enabled by default. The exposed fields are configured through the metadata manager.
Dataverse https://guides.dataverse.org/en/latest/admin/harvestserver.html	From the 'Harvesting Server' section in the Dashboard, the OAI-PMH service is enabled, and sets are defined. The endpoint is usually /oai.

To ensure the proper functioning of the OAI-PMH protocol, there are tools known as validators that evaluate the protocol's operability. One of the most well-known tools is the OAI-PMH Validator. This validator checks whether sets are adequately defined and whether records correctly export fields such as dates, identifiers, types, language, and other relevant information. It also allows for reviewing XML responses to detect errors or poorly structured formats. To analyze a repository's protocol functionality, simply enter the base URL, and the system will perform the analysis.

Figure 2

OAI-PMH Validator.



To ensure the proper functioning of the protocol, it is necessary to carry out maintenance activities and periodically monitor its status. Additionally, it is essential to update metadata mappings when recommendations from COAR, OpenAIRE, or other aggregators are revised. For ongoing monitoring, it is also essential to document the endpoints[2] and maintain coordination with other systems that depend on the repository.

4. Adoption of persistent identifiers

A persistent identifier is a unique, durable, and resolvable digital reference to a specific object, such as an article, dataset, software, person, or organization. These identifiers are designed to remain valid and accessible over time, even if the object's physical location or hosting server changes (Meadows et al., 2019).

Typically, a persistent identifier has three essential components:

- Global uniqueness, which means it includes a controlled syntax and a namespace governed by clearly defined authorities;
- Persistence, which ensures stable links and resolution functions, as well as persistent schemas and referenced objects; and
- Resolvable for both humans and machines, providing information on how to find, access, or use the referenced object (De Castro et al., 2023).

Persistent identifiers (PIDs) are important for optimizing the visibility and citability of publications, as they make it easier for search engines, academic repositories, and analytics

2 An endpoint is a specific address or URL through which an external system can interact with the repository to access its services or data.

tools to automatically find and link to documents without relying on unstable URLs. They also provide long-term stability by promoting resource accessibility, as they combat "link rot"[3] and "content drift"[4]: even if the object is moved, its persistent identifier will still resolve correctly. Finally, when PIDs are associated with structured metadata and supported by robust infrastructures, they enhance the trustworthiness and reputation of the content.

Among the most widely recognized PIDs is the Digital Object Identifier (DOI). It is the most used identifier for articles, books, datasets, and software. This system combines a permanent identifier with mandatory metadata and guaranteed resolution. When a DOI is resolved, it leads to a landing page with metadata, enhancing visibility and citation tracking. DOIs are assigned by registration agencies such as CrossRef and DataCite and are generally more costly than other identifiers.

Another option is the Handle system, a non-commercial identifier that has been used since 1995. Its main goal is to provide persistent identification and resolution services, operated centrally by the Corporation for National Research Initiatives (CNRI). A Handle identifier consists of a prefix that identifies the authority, along with a suffix that refers to the object being identified. Handle is the technical foundation of DOI and is more affordable to implement. Some systems, such as DSpace, integrate the Handle system by default; however, it must be acquired and configured to resolve resources through the identifier properly (https://wiki.lyrasis.org/display/DSDOC8x/Handle.Net+Registry+Support).

An alternative is the ARK (Archival Resource Key) persistent identifier, a system designed to provide durable and reliable links to digital objects, particularly useful in libraries, archives, and museums. Unlike other identifiers such as DOI or Handle, ARK is cheaper, decentralized, and highly flexible (https://arks.org), allowing institutions to generate and manage their identifiers without relying on a central registration authority. Its typical format is ark:/NAAN/identifier, where the NAAN identifies the issuing organization. A distinctive feature of ARK is its ability to provide access not only to the digital object but also to its metadata and a persistence commitment statement, which reinforces transparency and trust in long-term preservation. This system has been widely adopted by institutions such as the U.S. Library of Congress and the California Digital Library, supporting the visibility and traceability of cultural and academic resources.

Table 3
Comparison of persistent identifiers.

	DOI	Handle	ARK
Management	Centralized (DataCite, Crossref)	Distributed (CNRI)	Decentralized (institutional)
Structure	10.1234/abc123	20.5000/xyz456	ark:/12345/x6789
Resolution	Yes, via https://doi.org/	Yes, via https://hdl.handle.net/	Yes, via https://n2t.net/ or locally
Metadata Access	Yes (mandatory landing page)	Yes (depending on usage)	Yes (via inflection, i.e., a modifier character in the URL)
Guaranteed Persistence	High (by contract)	High (depends on repository)	Variable (based on institutional policy)

3 Link rot occurs when a hyperlink no longer leads to the intended content because the page has been moved, deleted, or the domain is no longer active.

4 Content drift occurs when the content at a given URL changes over time, so it no longer reflects what was originally cited or intended, even though the link still works.

	DOI	Handle	ARK
Cost	Requires paid membership and may include additional cost per DOI assignment	Annual fee of USD 50	No payment or membership required
Typical Uses	Articles, datasets, software	Repository objects (DSpace, Fedora)	Archives, libraries, digital museums

To implement a persistent identifier, the first step is to select which one will be used. It is worth noting that identifiers are not mutually exclusive and can be combined. For example, a resource can have both a DOI and an ARK.

It is necessary to register with a persistent identifier provider DOIs, this can be obtained through one of the registration agencies, such as Crossref or DataCite, although both require a paid membership and may involve a fee for minting each DOI.

The Handle system can be acquired through CNRI by paying USD 50 and linking it to a compatible repository system (https://www.handle.net/payment.html). Once configured, each deposited object receives an identifier with an authorized prefix, assigned by CNRI, and a unique suffix. These identifiers are resolved via https://hdl.handle.net/, ensuring long-term accessibility even if the resource's physical location changes.

ARK can be obtained by requesting a Name Assigning Authority Number (NAAN) at arks.org (https://arks.org/about/getting-started-implementing-arks). A NAAN is a unique number that identifies the ARK-issuing institution within the system. It functions as an official prefix that ensures each organization creating ARK identifiers has its exclusive namespace.

Afterward, it is necessary to configure the repository to issue and maintain persistent identifiers. This will depend on the specific software used to manage the institutional repository and can be done through registration agencies, the use of plugins, integration within the system itself, or by managing key configuration files.

Finally, the assignment of PIDs must be integrated into the workflow for setting up new resources, specifically the "identifier" field. It is also essential to ensure that the identifier correctly resolves to a landing page for the object, including its metadata and access to the resource.

PIDs such as DOI, Handle, and ARK are fundamental tools for strengthening the visibility of documents in a digital repository. By providing stable, unique, and long-lasting links, they ensure that resources remain easily findable, accessible, and citable, even when their technical location changes over time. Moreover, by being integrated into global resolution infrastructures and associated with structured metadata, these identifiers facilitate discovery by search engines, harvesters, academic citation systems, and open data networks. Altogether, PIDs ensure that documents are not only preserved but also disseminated and recognized in today's digital environments.

5. SEO Optimization of the repository

Search Engine Optimization (SEO) in digital repositories is a key strategy for increasing the visibility, accessibility, and impact of the academic, scientific, and cultural content they host. Despite having structured metadata and preservation standards, many repositories fail to rank well in search engines like Google or Bing, limiting the organic discovery of their resources by

users.

Implementing SEO best practices, including proper use of HTML tags, exposing Dublin Core metadata in schema.org format, creating user-friendly URLs, generating sitemaps, and enabling automatic indexing, improves how search engines interpret content. These measures facilitate accurate understanding, categorization, and prioritization of documents.

Moreover, combining SEO with persistent identifiers, such as DOI, Handle or ARK, reinforces the stability and traceability of resources on the web.

In a digital environment where attention is limited and competition for visibility is high, optimizing a repository's SEO is not just a technical improvement, but a strategic action to ensure that deposited resources fulfill their mission of being found, used, and cited.

The following elements are recommended for optimizing SEO in a repository:

5.1. Proper use of semantic HTML tagging

Use semantic tags such as <title>, <meta name="description">, <meta name="citation_doi">, <meta name="citation_author">, <h1>, <h2>, <article>, <section>, among others, to help search engines understand the structure of the content. This also helps tools like Altmetric better track the metrics of a particular resource (https://help.altmetric.com/support/solutions/articles/6000240582-required-metadata-for-content-tracking) (Reyes-Lillo & Pastor-Ramon, 2024).

It is crucial to ensure that each resource page (such as a document) has a unique and descriptive <title>. Additionally, including enriched metadata using schema.org or Dublin Core embedded in <meta> tags or JSON-LD format is recommended.

Below, you can see an example of proper semantic tagging using Dublin Core:

Figure 3

Example of Dublin Core embedded in <meta> tags.

```
<head>
  <meta name="DC.title" content="How to improve your metadata" />
  <meta name="DC.creator" content="Laura Martinez" />
  <meta name="DC.date" content="2023-11-15" />
  <meta name="DC.identifier" content="http://doi.org/10.1234/test-doi.2025.001" />
  <meta name="DC.language" content="es" />
  <meta name="DC.type" content="Book" />
</head>
```

And an example of semantic tagging using JSON-LD:

Figure 4

JSON-LD Example.

```
<script type="application/ld+json">
  {
    "@context": "https://schema.org",
    "@type": "Book",
    "name": "How to improve your metadata",
    "author": {
      "@type": "Person",
      "name": "Laura Martinez"
    },
    "datePublished": "2023-11-15",
    "identifier": {
      "@type": "PropertyValue",
      "propertyID": "Handle",
      "value": "http://doi.org/10.1234/test-doi.2025.001"
    },
    "inLanguage": "es",
    "license": "https://creativecommons.org/licenses/by/4.0/",
    "publisher": {
      "@type": "Organization",
      "name": "Universidad Nacional"
    }
  }
  </script>
```

5.2. Creating and maintaining an XML sitemap

Creating and maintaining an XML sitemap in a digital repository is essential for enhancing the indexing and visibility of content to search engines like Google or Bing. A sitemap acts as a structured map that lists all relevant pages of the repository, allowing search engines to discover new documents, updates, or deposited resources quickly.

For it to function correctly, the sitemap must include only public, permanent, and accessible URLs (such as those containing persistent identifiers like DOIs or Handles), and it should be updated automatically whenever content is added or modified. Additionally, it must be correctly referenced in the robots.txt file and submitted to tools like Google Search Console to maximize its effectiveness.

A well-implemented sitemap not only speeds up indexing but also improves the SEO ranking of resources, increasing their reach and usage within the academic and scientific ecosystem.

5.3. Proper configuration of the robots.txt file

The robots.txt file plays a key role in the SEO optimization of digital repositories, as it controls how search engines access, crawl, and index their contents. This file, located at the root of the website, tells robots (such as Googlebot or Bingbot) which parts of the repository can be explored and which should be excluded.

Proper configuration allows search engines to access the pages of digital objects, such as landing pages with persistent identifiers. On the other hand, it can block irrelevant or sensitive paths, such as administrative areas or navigation filters that could generate duplicate content. For this reason, the file must not block relevant paths linked to persistent identifiers, such as /handle/ or /ark:/.

A simple example is the following:

Figure 5

Example of elements to consider in the robots.txt file.

```
User-agent: *
Allow: /handle/
Disallow: /admin/
Disallow: /private/
Sitemap: https://repository.org/sitemap.xml
```

If not configured correctly, the robots.txt file can accidentally prevent the indexing of important resources, negatively affecting their visibility and discoverability in search results. Additionally, it should include a reference to the sitemap.xml file, making it easier for search engines to perform a structured crawl of the content.

5.4. Apply SSR (Server-Side Rendering) or hybrid rendering

Applying SSR (Server-Side Rendering) or a hybrid rendering approach in a digital repository is a technical strategy that helps improve the site's visibility and performance, especially in an increasingly AI-driven and automated indexing web environment.

Unlike CSR (Client-Side Rendering), where content is dynamically generated in the browser, SSR allows pages to be generated on the server before being sent to the user or search engine bot. This has multiple benefits: first, it improves SEO, as search engines can immediately access structured content without relying on JavaScript to render it. Second, it speeds up initial load times, enhancing user experience and supporting navigation from mobile devices or slow networks.

Moreover, by implementing hybrid rendering, SSR combined with CSR, an optimal balance is achieved between performance, interactivity, and visibility, making the repository effective for both humans and indexing bots or generative engines.

There are tools like Next.js, Nuxt, or Rendertron that help adapt sites to be SEO-friendly and compatible with these rendering strategies.

6. Generative Engine Optimization in repositories: a factor to consider

Generative Engine Optimization (GEO) is an emerging concept that refers to the optimization of digital content for generative search engines, such as ChatGPT, Google Search Generative Experience (SGE), or Perplexity, which use artificial intelligence (AI) to respond with directly generated text, rather than simply displaying links as traditional SEO does (Daniels, 2025).

GEO aims to adapt the way digital content is structured and tagged so that language models (LLMs) can correctly interpret it, reference it in their responses, and integrate it into AI-generated content (Aggarwal et al., 2024).

Just as SEO optimizes content to be more visible on Google, GEO optimizes content to be understood, cited, and used by AI-powered answer generation engines.

In this context, repositories can also take specific actions to optimize their content based on GEO strategies. For example, structuring metadata using technologies like schema.org, JSON-LD, or Open Graph is highly recommended to support the inclusion of their content in AI-generated responses.

Additionally, it is essential to provide clear and accessible content, avoiding hidden or overly technical language that may be difficult for LLMs to comprehend. It is also necessary to clearly indicate flexible intellectual property licenses, such as Creative Commons, to facilitate content reuse.

Moreover, the use of descriptive landing pages, persistent identifiers to ensure traceability, providing RDF or JSON files, and enabling API access are key factors that help improve content for processing by LLMs.

In summary, GEO represents a new visibility paradigm for digital repositories, where it is no longer enough to appear on Google; content must be structured, accessible, and understandable by language models.. Implementing GEO strategies not only increases the reach of resources but also prepares the repository to integrate into the AI-based search and discovery ecosystem that is shaping the future of knowledge access.

7. Conclusion

Visibility optimization in digital repositories requires a comprehensive approach that combines technical, regulatory, and strategic aspects. The quality of metadata ensures that resources are understandable and reusable by both humans and machines; interoperability allows for their seamless integration into global information networks; and persistent identifiers guarantee their traceability and long-term access.

At the same time, strong SEO optimization improves search engine ranking, while incorporating a GEO perspective expands the reach of content to generative artificial intelligence agents.

Together, these elements strengthen the visibility, impact, and circulation of the knowledge hosted in repositories, aligning them with the principles of open science and equitable access to information.

8. Funding

This work is part of the Project "Parameters and strategies to increase the relevance of media and digital communication in society: curation, visualisation and visibility (CUVICOM)". Grant PID2021-123579OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU.

9. References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). *GEO: Generative Engine Optimization* (arXiv:2311.09735). arXiv. https://doi.org/10.48550/arXiv.2311.09735

Chipangila, B., Liswaniso, E., Mawila, A., Mwanza, P., Nawila, D., M'sendo, R., Nyirenda, M., & Phiri, L. (2024). Controlled vocabularies in digital libraries: Challenges and solutions for increased discoverability of digital objects. *International Journal on Digital Libraries*, *25*(2), 139–155. https://doi.org/10.1007/s00799-023-00374-1

Daniels, C. (2025). *From SEO to GEO: How agencies are navigating LLM-driven search*. Campaign Asia. https://www.campaignasia.com/article/from-seo-to-geo-how-agencies-are-navigating-llm-driven-search/501593

De Castro, P., Herb, U., Rothfritz, L., & Schöpfel, J. (2023). *Building the plane as we fly it: The promise of Persistent Identifiers*. Zenodo. https://doi.org/10.5281/zenodo.7258286

Eells, L., Kelly, J., & Farrell, S. (2024). Repository (R)evolution: Metadata, interoperability, and sustainability. *Journal of Librarianship and Scholarly Communication*, *12*(1), Article 1. https://doi.org/10.31274/jlsc.16890

Ma, S., Lu, C., Lin, X., & Galloway, M. (2009). Evaluating the metadata quality of the IPL. *Proceedings of the American Society for Information Science and Technology*, *46*(1), 1–17. https://doi.org/10.1002/meet.2009.1450460249

Meadows, A., Haak, L. L., & Brown, J. (2019). Persistent identifiers: The building blocks of the research information infrastructure. *Insights*, *32*(1). https://doi.org/10.1629/uksg.457

OpenAIRE. (2020). OpenAIRE Guidelines—OpenAIRE Guidelines documentation. Retrieved from: *https://guidelines.openaire.eu/en/latest*

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. https://doi.org/10.7717/peerj.4375

Reyes-Lillo, D., Morales-Vargas, A. & Rovira, C. (2025). Visibility, discoverability, findability, Search Engine Optimization (SEO) and Academic SEO in digital repositories: A scoping review. *BiD: textos universitaris de biblioteconomia i documentació*, (54). https://doi.org/10.1344/BID2025.54.06

Reyes-Lillo, D., & Pastor-Ramon, E. (2024). Use of handle in institutional repositories and its relationship with alternative metrics: A case study in Spanish-speaking America. *Hipertext. Net*, *29*, Article 29. https://doi.org/10.31009/hipertext.net.2024.i29.17

Swan, A. (2010, February). *The open access citation advantage: Studies and results to date*. https://eprints.soton.ac.uk/268516/

UNESCO. (2021). *UNESCO recommendation on open science*. https://unesdoc.unesco.org/ark:/48223/pf0000379949

Van Hooland, S., & Verborgh, R. (2015). *Linked data for libraries, archives and museums: How to clean, link and publish your metadata*. Facet. https://doi.org/10.29085/9781783300389

Westbrook, R. N., Johnson, D., Carter, K., & Lockwood, A. (2012). Metadata clean sweep: A digital library audit project. *D-Lib magazine*, *18*(5/6). https://doi.org/10.1045/may2012-westbrook