Google Scholar Indexing for Institutional Repositories

Presentation for Couperin consortium
October 2021

Monica Westin, Google Scholar
Overview

1. How Google Scholar indexes repositories
2. Collaboration among HAL and Google Scholar
3. The crucial role of accurate metadata for indexing
4. How to ensure best possible metadata on the HAL platform
5. How to ensure best possible metadata on the DSpace platform
6. How to check your repository’s coverage in Scholar
7. Additional guidelines and resources
8. Questions and discussion
How Google Scholar indexes repositories
How the Google Scholar indexing system works

- The Scholar indexing system crawls the entire web looking for scholarly publications: articles, books, reports, theses, conference proceedings, preprints ...
- The indexing system identifies scholarly content, determines each item’s bibliographic metadata, and groups all versions of an item together with this metadata in search results
We try to make repositories visible worldwide

- For uniquely held items (e.g. dissertations), repository content is the primary link

THE MITOCHONDRIAL NATURE OF THE DNA REPAIR PROTEIN APE1
A Barbiasi - 2018 - iria.sissa.it

Mitochondria arose around two billion years ago from the engulfment of an α-proteobacterium by a precursor of the modern eukaryotic cell. Nowadays they are the principal source of energy production and metabolism that are fundamental processes for the survival and wellbeing of the cell. Mitochondria are organelles with a double membrane, the outer (OM) and the inner membrane (IM), and together they delimit two aqueous compartments: the matrix, in the inner side, and the intermembrane space (IMS)(Figure 1).

🌟 99 Related articles  All 3 versions  ⬤

Showing the best result for this search. See all results
For formally published articles, repository content appears as the access link and/or in “All XX versions”
Directed complete poset models of $T_1$ spaces

D Zhao, X Xi - ... Proceedings of the Cambridge Philosophical Society, 2018 - cambridge.org

A poset model of a topological space $X$ is a poset $P$ such that the subspace $\text{Max}(P)$ of the Scott space $\mathcal{E}P$ is homeomorphic to $X$, where $\text{Max}(P)$ is the set of all maximal points of $P$. Every $T_1$ space has a (bounded complete algebraic) poset model. It was, however, not ...
What Scholar needs for indexing repository publications

- Access to crawl the site
- Way to find all urls for articles—usually sitemap for repositories
- Bibliographic information in the form of machine-readable metadata tags ("metatags"), on by default for DSpace repositories past version 1.7

Bibliographic metatags tell the Scholar indexing system what the metadata for an article is: title, author, publication date, etc.

"Citation_pdf_url" metatag tells the indexing system which file to associate with this metadata

```
<meta name="citation_title" content="The testis isoform of the phospholysase kinase catalytic subunit (Phx-T) plays a critical role in regulation of glycozen mobilization in developing lung">
<meta name="citation_author" content="Liu, Li">
<meta name="citation_author" content="Rannels, Stephen R.">
<meta name="citation_author" content="Falconieri, Mary">
<meta name="citation_author" content="Phillips, Karen S.">
<meta name="citation_author" content="Wolpert, Ellen B.">
<meta name="citation_author" content="Weaver, Timothy E.">
<meta name="citation_publication_date" content="1996/05/17">
<meta name="citation_journal_title" content="Journal of Biological Chemistry">
<meta name="citation_volume" content="271">
<meta name="citation_issue" content="20">
<meta name="citation_firstpage" content="11761">
<meta name="citation_lastpage" content="11766">
<meta name="citation_pdf_url" content="http://www.example.com/content/271/20/11761.full.pdf">```
View source code from repository item page to view metatags

Right click or keyboard command to “View Page Source,” depending on your browser

Search HTML source for “citation_” to view metatags
Collaboration among HAL and Google Scholar - and role of repository administrators
HAL & Google Scholar teams have long collaborated

- HAL and Google Scholar teams have been working closely for many years
- We collaborate closely on indexing related features & infrastructure
- This collaboration has been very successful in ensuring broad coverage of French repositories, increasing global visibility for French research outputs
- We continue to work together to ensure repositories are indexed comprehensively and accurately as scale of repository content grows exponentially
Role of repository administrators

- Accurate & effective indexing requires both platform and repository level effort

- In the rest of the slides, we present what repository administrators can do to make sure hosted articles are indexed well
  
  - Focus on automated, sustainable approaches rather than manual fixes, which are increasingly difficult at the scale of most current institutional repositories
The crucial role of accurate metadata for indexing.
Inaccurate metatags prevent indexing

- Indexing in Scholar requires accurate metatags
  - e.g. the order of author metatags in the HTML source controls the author order that appears in Scholar search results.

- Incorrect metadata results in unhappy authors & items not ranked as they should be due to missing citations.

- The indexing system automatically detects sites with frequent metadata errors and stops including them.
Metatags need to match the version of record:

PSICOLOGIA, SAÚDE & DOENÇAS 2009, 10 (2), 217-225

ADAPTAÇÃO E ESTUDO DA ESCALA DE VALOR DA SAÚDE

Filipa Pimenta¹, Isabel Leal² & João Maroco¹

¹Unitade de investigação em Psicologia e Saúde, ID
²ISPA

RESUMO: O presente estudo pretende averiguar se pessoas com diferentes comportamentos de consumo de tabaco valorizam a sua saúde de forma discrepante, é igualmente objectivo desta investigação adaptar a Escala de Valor da Saúde (Lau, Hartman, & Ware, 1986) à população Portuguesa. Para tal aplicou-se o instrumento a uma amostra de 380 estudantes universitários (fumadores regulares, ocasionais e ex-fumadores). Os resultados demonstram que pessoas com diferentes padrões de consumo tabágico atribuem um valor elevado à sua saúde, não existindo diferenças significativas entre os três grupos (F=1,594; p=0,205). O estudo das características psicométricas do
Common metatag errors for repository publications

- Incorrect date listed in publication date metatags, most commonly listing the online/upload date as the publication date
- Missing authors in citation_author metatags, most commonly only including authors from institution or only listing the first author of a publication
- Incorrect author order, most commonly adding author(s) from the institution first
How to ensure best possible metadata on the HAL platform
HAL metadata best practices begins with deposit

- Metadata automation at deposit helps avoid common metadata errors
- Add DOI at upload whenever available
  - Adding DOI initially retrieves metadata from Crossref for improved accuracy
- When DOI doesn’t exist, metadata can be extracted from uploaded PDF directly
The type of your document is 'Journal articles'.

In case of error, you can choose the document type in the section "Complete the document metadata."

Nous avons complété votre dépôt à partir de l'identifiant:

10.1080/15362426.2017.1384769

Fulltext language, Title, Record date, Journal, Volume, Page number, Number, Identifiers

Aristotle's Rhetorical Energia : An Extended Note
Metadata: what HAL repository administrators can do

- Reviewing metadata for possible errors, including any occasional errors caused by automated metadata extraction from PDFs
- Ensuring accurate metadata for items without DOIs, especially where library is hosting the version of record, e.g. theses and dissertations
How to ensure best possible metadata on the DSpace platform
Fixing common DSpace metadata issues

- Several common errors can often be fixed systematically via code patches or platform upgrades
  - Online/upload date listed as publication date in metatags
  - Incorrect author order in author metatags
- Other common metadata errors currently require manual corrections, e.g. missing authors in author metatags
Incorrect publication dates in citation_date metatag

- Test by comparing date citation_date tag with date in the version of record. If date listed in citation_date tag is later than the version of record, likely online date is being used as publication date.

- Systematic fix: patch for DSpace repositories available at https://github.com/DSpace/DSpace/pull/2294.patch
  - More information here: https://jira.duraspace.org/browse/DS-4104
Incorrect author order in citation_author metatags

- Systematic fix: If using DSpace versions 5.0, 5.1, 5.2, and 5.3, patch for DSpace repositories available at https://github.com/DSpace/DSpace/pull/999
  - More information here: https://jira.duraspace.org/browse/DS-2679
  - Upgrading to DSpace v5.4 or above will also fix this problem

- Manual fix: Adjust author metatag order to match author order in the version of record
Checking your repository’s coverage
How to do a Scholar coverage check for your repository:

What doesn’t work

- The result count of searching your repository site ("site:XXX") is not an accurate indicator of Scholar coverage.
- This is because the number listed in Scholar search results for a site only applies to the primary links — and as described earlier, the repository content is often not the primary link (in “All XX versions”).
How to do a coverage check for your repository: what does work

- Search in Scholar for titles of several dozen randomly selected items across the repository and see if these papers are included

- Be sure to check the “All XXX versions” link as well, as often the repository version will not be the primary link
Additional Scholar indexing guidelines and resources
Google Scholar guidelines and resources for repositories

1. Google Scholar inclusion guidelines & troubleshooting guidelines


Final thoughts

- Repositories are a crucial pathway for scholarly access
  - But what can’t be discovered can’t be built upon

- Platform automation for accurate metadata is key for sustainable, scalable repositories

- We need to work together to help researchers everywhere advance from a shared frontier of knowledge
Thank you for joining us!
Questions?

mwestin@google.com