Google Scholar Indexing for DSpace Repositories

Monica Westin, Google Scholar
mwestin@google.com
Overview

1. How the Google Scholar indexing system works
2. Common repository indexing problems
3. Suggested fixes
4. How to check your repository’s coverage in Scholar
5. Resources and troubleshooting guidelines
6. Questions
How the Google Scholar indexing system works

- Google Scholar 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
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 $\text{EP}$ 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 ...

Directed complete poset models of $T_1$ spaces

D Zhao, X Xi - 2018 - repository.nie.edu.sg

A poset model of a topological space $X$ is a poset $P$ such that the subspace $\text{Max}(P)$ of the Scott space $P$ 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 ...

Directed complete poset models of $T_1$ spaces

D Zhao, X Xi - Mathematical Proceedings of the Cambridge ..., 2018 - adsabs.harvard.edu

Directed complete poset models of $T_1$ spaces

D ZHAO, X XI - Mathematical Proceedings of the Cambridge ..., 2018 - search.proquest.com

A poset model of a topological space $X$ is a poset $P$ such that the subspace $\text{Max}(P)$ of the Scott space $\{P\}$ 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

- Access to crawl the site
- Way to find all urls for articles-- sitemap or browse by date
- 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
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
Metatags should match the version of record:

ADAPTAÇÃO E ESTUDO DA ESCALA DE VALOR DA SAÚDE

Filipa Pimenta1, Isabel Leal12 & João Maroco11

1Unidade de investigação em Psicologia e Saúde, I&D
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 tabagico 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
Repository indexing errors

- Commonly: Incorrect bibliographic metatags
  - Sites with widespread metadata errors can’t be indexed

- Occasionally: Site outages that occur while the indexing system is looking for publications in your repository

- Occasionally: crawler issues, including blocking the Googlebot crawler, slow responsiveness or errors to crawlers, or limiting the crawl speed.
Problem: Incorrect publication dates in citation_date metatag

- Occurs most commonly when upload/online date is given as publication date when no publication date exists, often via batch uploads, e.g.
  - `<meta name="citation_date" content="2014"/>

versus publication date in version of record:

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

ADAPTAÇÃO E ESTUDO DA ESCALA DE VALOR DA SAÚDE

Filipa Pimenta¹, Isabel Leal¹,² & João Maroco¹,²
¹Unidade de Investigação em Psicologia e Saúde, I&D
²ISPA
Fix: 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.

- 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
Problem: Incorrect author order in citation_author metatags

- Occurs when the author metatags are listed out of order in the source code, e.g.
  - `<meta name="citation_author" content="Leal, Isabel"/>
  - `<meta name="citation_author" content="Pimenta, Filipa"/>
  - `<meta name="citation_author" content="Maroco, João"/>

versus author order in version of record:

Filipa Pimenta¹, Isabel Leal¹,² & João Maroco¹,²
Fix: Incorrect author order in citation_author metatags

- 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

- Fix: Adjust author metatag order to match author order in the version of record
  - <meta name="citation_author" content="Pimenta, Filipa" />
  - <meta name="citation_author" content="Leal, Isabel" />
  - <meta name="citation_author" content="Maroco, João" />

Filipa Pimenta¹, Isabel Leal¹,² & João Maroco¹,²
Problem: Missing authors in metatags

- Most commonly only including authors from institution or only listing the first author of a publication, e.g.
  - <meta name="citation_author" content="Leal, Isabel"/>

  vs. Filipa Pimenta¹, Isabel Leal¹² & João Maroco¹²

- Fix: Include all authors listed in the version of record in author metatags, not just authors from your own institution. (And again, list authors in the order that they appear in the published version.)
Problem: Too many authors in metatags

- Most commonly listing the advisor as an author for a thesis/dissertation, e.g.
  - <meta name="citation_title" content="Jungian approaches to underwater basketweaving"/>
  - <meta name="citation_author" content="Sara Student"/>
  - <meta name="citation_author" content="Professor Patricia"/>

- Fix: Only include one author of a thesis or dissertation in author metatag—the student who wrote it
  - <meta name="citation_title" content="Jungian approaches to underwater basketweaving"/>
  - <meta name="citation_author" content="Sara Student"/>
Problem: Trailing information in metatags

- Most commonly repository name or article type appended to title metatag, e.g.
  - <meta name="citation_title" content="Jungian approaches to underwater basketweaving Northern California College Repository"/>
  - <meta name="citation_title" content="Jungian approaches to underwater basketweaving Thesis"/>

- Fix: (1) Avoid adding any elements other than article bibliographic information, (2) Remove non-bibliographic information from metatags
  - <meta name="citation_title" content="Jungian approaches to underwater basketweaving"/>
Problem: Multilinguals in metatags

- Combining different scripts in metatags results in mixed bibliographic information that prevents the item from being ranked in Scholar search results — and is misleading for users

- ex: including the translated version of the title in title metatags, e.g.
  - <meta name="citation_title" content="War and Peace == Война и мир" />

- ex: listing authors in native script of home institution when it is not the language in which the article was written, e.g.
  - <meta name="citation_author" content="Толстой, Лев Николаевич " />
  - <meta name="citation_author" content="Tolstoy, Lev Nikolayevich "/>
Fix: Multilinguals in metatags

- Fix: use the language of the full text/abstract for all metatags. Don’t duplicate metatag information in multiple scripts

Л. Толстой. «Война и мир».

Часть 1

Прочтите приведённый ниже фрагмент текста и выполните задания B1-B7; C1-C2.

Вследствие этого страшного гула, шума, потребности внимания и деятельности, Тушин не испытывал ни малейшего неприятного чувства страха, и мысль, что его могут убить или больно ранить, не приходила ему в голову. Напротив, ему становилось все веселее и веселее. Ему казалось, что уже очень давно, едва ли не вчера, была та минута, когда он увидел неприятеля и сделал первый выстрел, и что ключок поля, на котором он стоял, был ему давно знакомым, родственным местом. Несмотря на то, что он все помнил, все соображал, все делал, что мог делать самый лучший офицер в его положении, он находился в состоянии, похожем на лихорадочный бред или на состояние пьяного человека.

Из-за оглушающих со всех сторон звуков своих орудий, из-за свиста и ударов снарядов неприятеля, из-за вида вспотевшей, раскрасневшейся, горячей, бежавшей около орудий прислуги, из-за вида крови людей и лошадей, из-за вида дымок неприятеля на той стороне (после которых всякий раз прилетало ядро и било в землю, в человека, в орудие или в лошадь), — из-за вида этих предметов у него в голове установился свой фантастический мир, который составлял его наслаждение в эту минуту.

<meta name="citation_title" content="Война и мир" />

<meta name="citation_author" content="Толстой, Лев Николаевич" />
**Problem:** Site outages

- Outages result in dead pages in search results
- Extended or repeated site outages obstruct the indexing system’s crawler from finding the items in your repository
- Extended outages can also result in the site dropping out of the index
- **Fix:** Don’t keep your repository site down for extended periods of time, e.g. multiple days
- **Fix:** If you are migrating to a new site, keep the old site functional while developing the new version. When the new version is ready, change DNS lookup from the old server to the new one.
Example: site outage for large DSpace repository in Europe

Outage occurs
Items are re-added to Scholar index after site is restored
Problem: invalid HTTPS certificate

- When the https version of a site returns errors due to an invalid certificate, the indexing system is blocked from crawling the site and is forced to removed it from the Scholar index.

- **Fix:** To test, open an article in several different browsers (Chrome, Safari, etc). If you see warnings that the connection is not private, contact your SSL certificate provider to fix.
Other repository errors

- Occasionally crawler access is blocked by other site configurations
  - **Fix:** Keep default crawler settings for DSpace sites

- Repository moved without redirects, or items renumbered without redirects.
  - **Fix:** Set up redirects whenever URLs for publications in the repository change.

- Adding an interstitial for accessing fulltext
  - **Fix:** Skip interstitials for users clicking on search results

- Adding a cover page to fulltext articles
  - **Fix:** Host PDFs articles as-is
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”)

Not effective!
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
Google Scholar guidelines and resources for repositories

1. Google Scholar inclusion guidelines & troubleshooting guidelines


3. Nigerian DSpace User Group
Thank you for joining us! Questions?