Google Scholar Indexing for Open Journal Systems (OJS)

Monica Westin, Google Scholar
Overview

1. How the Google Scholar indexing system works
2. Common OJS indexing problems and suggested fixes
3. Best practices for OJS journal indexing
4. Google Scholar indexing guidelines and resources for OJS
5. 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
What Scholar needs for indexing

- URLs for all articles
- Bibliographic information in the form of machine-readable metadata tags ("metatags")

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 article landing 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

```html
<meta name="citation_journal_title" content="Ateliê Geográfico"/>
<meta name="citation_issn" content="1982-1956"/>
<meta name="citation_author" content="Elaiz Aparecida Mensch Buffon"/>
```
Metatags should match the published PDF:

- Title, author, and publication dates match
- Bibliographic metatags written in (only) language/script of the article
  - e.g. citation_title gives article title in Portuguese
Publication date metatag should match date of formal publication for the issue
OJS indexing errors

- Commonly: Incorrect or missing bibliographic metatags
  - Indexing system forced to skip sites with widespread errors

- Site outages that occur while the indexing system is looking for publications

- 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

- Known issue for OJS versions 3.0 and 2.4
- Usually upload date of article is included as publication date, e.g.
  - `<meta name="citation_date" content="2019"/>

versus issue date and date listed in PDF:
Fix: Incorrect publication dates in citation_publication_date metatag

- Patches for supported versions of OJS include:
  - OJS 3.0.1 and 3.0.2:
    https://github.com/pkp/ojs/commit/9105b7ebc397f3647d500d62d30654b56a4f1e60.diff
  - OJS 3.0.0:
    https://github.com/asmecher/ojs/commit/c46a9d7a0873ba21ab7fec1cd1453cae215f9a3f.diff
  - OJS 2.4.x:
    https://github.com/pkp/ojs/commit/6cf0078c0ff0556231696cf9979377381d18d721d.diff
- More information here: https://github.com/pkp/pkp-lib/issues/2739
Problem: Metatags no longer included after OJS upgrade

- Known issue for journals upgrading from OJS 2.x to OJS 3.0.1, 3.0.2, 3.1.0, and 3.1.1.
- In these cases, the Google Scholar plugin was not automatically enabled.
- To test, view the source code for a few articles in each journal that upgraded.
  - If there is no citation_title tag, your site is affected.
Fix: Metatags no longer included after OJS upgrade

- **Fix**: re-enable Google Scholar plugin manually via admin dashboard for OJS instances with a small number of journals
- **Fix**: upgrade to OJS 3.1.2
- **Fix**: use an SQL command for large OJS instances with many journals:
  
  https://github.com/pkp/ojs/blob/stable-3_1_2/dbscripts/xml/upgrade/3.1.2_update.xml#L41..L42
Fix: Select the “Google Scholar Indexing Plugin” checkbox
**Problem:** Language mix up for multilingual metatags

- Combining different languages/scripts in metatags results in mixed, duplicate bibliographic information that prevents the item from being indexed properly in Scholar search results — and is confusing 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/language 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 "/>
Problem: Using English by default for all metatags

- Using English in metatags, e.g. in the title metatags, when the article is in a different language results in citations to articles being missed and articles not being ranked as they should be.

ex: `<meta name="citation_title" content="Socio-environmental risk to human leptospirosis in the Metropolitan Urban Parcel of Curitiba, Paraná-Brazil: a methodological proposal based on GIS"/>

Risco socioambiental à leptospirose humana no Aglomerado Urbano Metropolitano de Curitiba, Paraná-Brasil: uma proposta metodológica apoiada em SIG

Social-Environmental risk to human leptospirosis in Urban Agglomeration Metropolitan of Curitiba, Paraná-Brazil: a methodological proposal supported by GIS

Elaiz Aparecida Mensch Buffon
Universidade Federal do Paraná
eambuffon@gmail.com

Francisco de Assis Mendonça
Universidade Federal do Paraná
chico@ufpr.br

Resumo
O conhecimento das áreas de risco socioambiental à leptospirose humana pode auxiliar no processo de tomada de decisão por gestores. A quantificação e o mapeamento desse risco por meio de um índice sintético possibilitam uma análise integrada da problemática. Nesse sentido, a presente pesquisa tem como objetivo propor e avaliar uma metodologia para o mapeamento do risco socioambiental à leptospirose humana por meio de um índice sintético que considera as vulnerabilidades e a exposição à doença. A proposta foi aplicada no Aglomerado Urbano Metropolitano de Curitiba, visando apresentar uma análise no nível de desagregação intrametropolitana e intrametropolitana, considerando um cenário de 2007 a 2013. Foram utilizadas técnicas estatísticas associada a cartografia de síntese (análise multiciúfeteria, cálculo da álgebra de mapas, cálculo de incidência e matriz de cruzamentos), que culminaram nos seguintes resultados: 1) a metodologia se mostrou eficaz para sintetizar um grande número de dados, que possibilitou reduzir fenômenos complexos de inter-relações em mensagens simples e, 2) os maiores índices de risco socioambiental estão localizados tanto em regiões periféricas,
Fix: Always use language of article for metadata

- Use only the language/script of the published PDF for all metatags. Don’t duplicate metatag information in multiple scripts.

ex: <meta name="citation_title" content="Risco socioambiental à leptospirose humana no Aglomerado Urbano Metropolitano de Curitiba, Paraná-Brasil: uma proposta metodológica apoiada em SIG" />
**Problem:** Badly formatted author names

- Common examples include incorrect “first name, last name” format, incomplete author names, and errors in spelling and capitalization, e.g.
  - `<meta name="citation_author" content="Filipa, Pimenta" />`
  - `<meta name="citation_author" content="Leal" />`
  - `<meta name="citation_author" content="Joao maoco" />`

  vs. author names listed in published PDF:

  Filipa Pimenta¹, Isabel Leal¹,² & João Maroco¹,²

- **Fix:** List complete author names in citation_author tags as they are written in the published PDF. Use either “last name, first name” or “first name last name” format in metatags.
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.
**Problem:** hacked OJS sites

- Large number of OJS sites have been hacked
  - Close to 150 hacked sites identified by indexing system
  - Hacked sites are used for commercial spamming
  - Hackers replace OJS code to return spammy pages to Google crawler
- Scholar indexing tries to identify hacked OJS sites
  - Hacked OJS sites are removed from the index
Fix: hacked OJS sites

- **Test:** fetch an article URL emulating the crawler

  ```bash
  curl -A Googlebot URL_ON_YOUR_SITE
  ```

  - Check a broad sample of randomly selected articles to see if they return a different page from what you see in the browser, or redirect to another site.

- Alert your hosting provider

- Alert the PKP community forum at [https://forum.pkp.sfu.ca](https://forum.pkp.sfu.ca)
Best practices for OJS journal indexing

- **Avoid customizing URLs**, which makes it more difficult to identify a journal site, and as a result takes longer to index
  - There is no need to register your site for it to be indexed. The Scholar crawler will automatically find the site.
- The standard OJS URL structure has worked well for a long time
  
  Example of standard URL:

Best practices for OJS journal indexing

- **Set up article-level HTTP 301 (permanent) redirects** whenever URLs for publications in the site change, e.g. when journal site moves or items are renumbered

- Example:

  http://ojs.statsbiblioteket.dk/index.php/her/article/view/25673

  now redirects to

  https://tidsskrift.dk/index.php/her/article/view/25673
Best practices for OJS journal indexing

- **Avoid outages during migrations**
  - 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
Best practices for OJS journal indexing

- Avoid repeated site outages
  - Outages result in dead pages in search results
  - Repeated outages obstruct the Scholar crawler from finding your publications:

<table>
<thead>
<tr>
<th>Date</th>
<th>06-20</th>
<th>06-21</th>
<th>06-22</th>
<th>06-23</th>
<th>06-24</th>
<th>06-25</th>
<th>06-26</th>
<th>06-27</th>
<th>06-28</th>
<th>06-29</th>
<th>06-30</th>
<th>07-01</th>
<th>07-02</th>
</tr>
</thead>
</table>
Best practices for OJS journal indexing

- **Avoid extended site outages**
  - Extended site outages obstruct the Scholar crawler from finding your publications
  - Extended outages can also result in the site dropping out of the index entirely:

<table>
<thead>
<tr>
<th>Date</th>
<th>06-20</th>
<th>06-21</th>
<th>06-22</th>
<th>06-23</th>
<th>06-24</th>
<th>06-25</th>
<th>06-26</th>
<th>06-27</th>
<th>06-28</th>
<th>06-29</th>
<th>06-30</th>
<th>07-01</th>
<th>07-02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>🟥</td>
<td>🟥</td>
<td>🟥</td>
<td>🟥</td>
<td>🟥</td>
<td>🟥</td>
<td>🟥</td>
<td>🟥</td>
<td>🟥</td>
<td>🟥</td>
</tr>
</tbody>
</table>

- **Don’t keep your OJS site down** for extended periods of time, e.g. multiple days
Google Scholar indexing guidelines and resources for OJS

1. Google Scholar inclusion & troubleshooting guidelines


2. “Indexing Repositories: Pitfalls & Best Practices” presentation from 2015 Open Repositories conference (targeted to repositories, but has good general guidelines for both Scholar & web indexing)


3. PKP Community Forum

https://forum.pkp.sfu.ca
Thank you for joining us!
Questions?

Contact me: mwestin@google.com