Exposing repository content to Google Scholar

Bram Luyten - @LuytenBram
Dec 12th 2018
Atmire Webinar
Overview

Basic principles

Check your repository with analyzer.atmire.com

Google Analytics

Most, if not all of the information presented today is present, in some shape or form on

But first
Does Google Scholar rely on OAI-PMH?
What does the "site:" operator tell you about coverage of your repository?
Does Google Scholar index all your items?
How long does it take for an item to get indexed?
If you don't see your item in the search results, does that mean it isn't indexed?
Is it required to register your repository somewhere in order to get it included?

https://www.google.com/support/scholar/bin/request.py
Site-wide principles

Google Scholar treats DSpace like any website
Robots.txt needs to be in the root of the domain
Robots.txt needs to reference a sitemap
Pages should load "fast enough"
It's alive!
How to test an item
Published peer review materials


Publishing peer review materials alongside research articles promises to make the peer review process more transparent as well as making it easier to recognize these contributions and give credit to peer reviewers. Traditionally, the peer review report, editor's letter and …

99 All 9 versions
Publishing peer review materials
Publishing peer review materials alongside research articles promises to make the peer review process more transparent as well as making it easier to recognize these contributions and give credit to peer reviewers. Traditionally, the peer review reports, editor's letters and ...
Item specific principles
Scholarly articles

"The content hosted on your website must consist primarily of scholarly articles - journal papers, conference papers, technical reports, or their drafts, dissertations, pre-prints, post-prints, or abstracts. Content such as news or magazine articles, book reviews, and editorials is not appropriate for Google Scholar. Documents larger than 5MB, such as books and long dissertations, should be uploaded to Google Book Search; Google Scholar automatically includes scholarly works from Google Book Search."
File format

"Your files need to be either in the **HTML** or in the **PDF format**. PDF files must have searchable text, i.e., you must be able to search for and find words in the document using Adobe Acrobat Reader.

Each file must **not exceed 5MB in size**. To index larger files, or to index scanned images of pages that require OCR, please upload them to Google Book Search."

Basic Principles - Item specific
The next slide is the most important slide in this webinar
Metadata Matters!

**Required fields**

- **citation_title**
- **citation_author** preferably one **author per tag** and **order correctly**
- **citation_publication_date** minimally year - use dc.date.issued
- **citation_pdf_url**


Basic Principles - Item specific
Author order pre-DSpace 5.4

Author order was systematically wrong in 5.0, 5.1, 5.2 and 5.3. This was resolved in DSpace 5.4 and as of DSpace 6

https://jira.duraspace.org/browse/DS-2679

If you're pre-DSpace 5.4, either patch or perform the minor update.
<meta name="citation_title" content="The testis isoform of the phosphorylase kinase catalytic subunit (PhK-T) plays a critical role in regulation of glycogen 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">
Metadata mapping configuration

```plaintext
google.citation_title = dc.title

google.citation_publisher = dc.publisher

google.citation_author = dc.author | dc.contributor.author | dc.creator

google.citation_date = dc.date.copyright | dc.date.issued | dc.date.available |
dc.date.accessioned

google.citation_language = dc.language.iso
...

https://github.com/DSpace/DSpace/blob/dspace-6_x/dspace/config/crosswalks/google-
metadata.properties

https://wiki.duraspace.org/display/DSDOC6x/Search+Engine+Optimization
```

Basic Principles - Item specific
Journal and Conference Papers

Optional fields

citation_journal_title, citation_conference_title
citation_issn
citation_isbn
citation_volume
citation_issue
citation_firstpage
citation_lastpage


Basic Principles - Item specific
Theses, Dissertations, Tech Reports

Optional fields

citation_dissertation_institution
citation_technical_report_institution for the name of the institution
citation_technical_report_number

Item principles RECAP

1. Scholarly content only

2. Google Scholar needs to be able to get to the full text, in HTML or PDF, max 5MB and clearly linked FROM the item page if it's not ON the item page itself.

3. The metadata on your item page needs to match with what's included in the full text file, watch especially out for correct authors and issue dates.

Basic Principles
Checking your repository and items

https://analyzer.atmire.com
To get your DSpace items and bitstreams indexed in Google Scholar, you need to follow two basic principles.

With this tool you can check if your repository already adheres to these principles.

This tool has been used to analyse 2447 repositories, containing a total of 30,191,556 items.
GET STARTED

Enter your repository URL here to initiate the analysis.

HTTPS://NEWDEMO.OPENREPOSITORY.COM
ANALYZE!
OPEN REPOSITORY HOME

RESULTS RETRIEVED ON FRI DEC 07 16:41:10 CET 2018

REFRESH THE RESULTS NOW

DSpace 5.7

XML Mirage 2 User Interface

Homepage loaded in 219 milliseconds

266 items

Items indexed in Google Scholar

Following items have been analysed with this tool. Items in red were not found in Google Scholar. For the other ones, at the very minimum the metadata was confirmed as indexed in Google Scholar. The file icon is shown in green when the full text was found in scholar, and red if it wasn't.
As you will see in the robots.txt analysis below, we didn't find any problems with your robots.txt file or sitemap. This is great. You can now proceed and test the indexing of individual items.

Enter an item url to start item analysis:

HTTPS://NEWDEMO.OPENREPOSITORY.COM/...

START ITEM ANALYSIS

ROBOTS.TXT AND SITEMAP REPORT

ROBOTS.TXT FILE FOUND
HTML MAP OR SITEMAP FOUND
HANDLE LINKS ALLOWED
BITSTREAM LINKS ALLOWED
BROWSE PAGES ALLOWED OR SITEMAP PRESENT
PLAN-S Compliant Test Item: Publishing peer review materials

Name: f1000research-7-1655-v1.xml
Size: 62.78Kb
Format: XML

Author
Beck, Jeffrey
Funk, Kathryn

URI
https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85056322635&origin=inward; http://hdl.handle.net/2384/58285

Abstract
Publishing peer review materials alongside research articles promises to make the peer review process more transparent and accessible.
PLAN-S COMPLIANT TEST ITEM: PUBLISHING PEER REVIEW MATERIALS

Beck, Jeffrey; Funk, Kathryn; Harrison, Melissa; McEntyre, Jo; Breen, Josie; Collings, Andy; Donohoe, Paul; Evans, Michael; Flintoft, Louisa; Hamelers, Audrey; Hurst, Phil; Lemberger, Thomas; Lin, Jennifer; O'Connor, Niamh; Parkin, Michael; Parker, Sam; Rodgers, Peter; Skipper, Magdalena; Stoner, Michael
This tool makes a query in Google Scholar to find your item. Unfortunately, Google is getting more and more aggressive in blocking robots and tools like this one to make automatic queries. So if our automated request has failed, you should check manually.

Click here to tryout the query that was used to detect your item. It is a simple query on the URL of the item. In case you don’t see a reference to your repository on the first page, make sure to follow the link "All X versions", and look for an entry about your repository there. Different occurrences of the same publication get clustered together.

If you have checked for yourself, and if you found the item, click here to indicate that the metadata for this item was FOUND in Google Scholar.
Following items have been analysed with this tool. Items in red were not found in Google Scholar. For the other ones, at the very minimum the metadata was confirmed as indexed in Google Scholar. The file icon is shown in green when the full text was found in scholar, and red if it wasn’t.

PLAN-S Compliant Test Item: Publishing peer review materials
Your Google Scholar traffic in Google Analytics
Identifying traffic from a specific source
<table>
<thead>
<tr>
<th>Source</th>
<th>Acquisition</th>
<th>Behaviour</th>
<th>Conversions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Users</td>
<td>New Users</td>
<td>Sessions</td>
</tr>
<tr>
<td>1. newdemo.openrepository.com</td>
<td>413 (62.29%)</td>
<td>409 (62.63%)</td>
<td>487 (59.32%)</td>
</tr>
<tr>
<td>2. newdemo.upgrade.openrepository.com</td>
<td>57 (0.80%)</td>
<td>57 (0.80%)</td>
<td>76 (9.26%)</td>
</tr>
<tr>
<td>3. 99-reasons-for-seo.com</td>
<td>35 (5.28%)</td>
<td>35 (5.35%)</td>
<td>35 (4.26%)</td>
</tr>
<tr>
<td>4. scholar.google.com</td>
<td>24 (3.62%)</td>
<td>24 (3.69%)</td>
<td>24 (2.92%)</td>
</tr>
<tr>
<td>5. t.co</td>
<td>19 (2.87%)</td>
<td>19 (2.91%)</td>
<td>19 (2.31%)</td>
</tr>
<tr>
<td>6. scholar.google.co.uk</td>
<td>18 (2.71%)</td>
<td>16 (2.45%)</td>
<td>18 (2.19%)</td>
</tr>
<tr>
<td>7. 9-reasons-for-seo.com</td>
<td>15 (2.26%)</td>
<td>15 (2.30%)</td>
<td>16 (1.95%)</td>
</tr>
<tr>
<td>8. newdemo.aws.openrepository.com</td>
<td>11 (1.66%)</td>
<td>11 (1.68%)</td>
<td>11 (1.34%)</td>
</tr>
<tr>
<td>9. google.com</td>
<td>7 (1.06%)</td>
<td>7 (1.07%)</td>
<td>7 (0.85%)</td>
</tr>
<tr>
<td>10. 17-reasons-for-seo.com</td>
<td>6 (0.90%)</td>
<td>6 (0.92%)</td>
<td>6 (0.73%)</td>
</tr>
<tr>
<td>Source</td>
<td>Acquisition</td>
<td>Behaviour</td>
<td>Conversions</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
|                 | Users       | New Users | Sessions    | Bounce Rate | Pages/Session | Avg. Session Duration | Bitstream downloads (Goal 1 Conversion Rate) | Bitstream downloads (Goal 1 Completions) | % of Total: 0.00%
<p>| scholar.google.com | 24 (29.27%) | 24 (30.00%) | 24 (29.27%) | 0.00% | 0.00 | &lt;00:00:01 | 0.00% | 0 (0.00%) |
| scholar.google.co.uk | 18 (21.95%) | 16 (20.00%) | 18 (21.95%) | 0.00% | 0.00 | 00:00:00 | 0.00% | 0 (0.00%) |
| scholar.google.co.id | 4 (4.88%)   | 4 (5.00%)  | 4 (4.88%)   | 0.00% | 0.00 | 00:00:00 | 0.00% | 0 (0.00%) |
| scholar.google.co.in | 4 (4.88%)   | 4 (5.00%)  | 4 (4.88%)   | 0.00% | 0.00 | 00:00:00 | 0.00% | 0 (0.00%) |
| scholar.google.com.au | 4 (4.88%)   | 4 (5.00%)  | 4 (4.88%)   | 0.00% | 0.00 | 00:00:07 | 0.00% | 0 (0.00%) |
| scholar.google.de  | 4 (4.88%)   | 4 (5.00%)  | 4 (4.88%)   | 0.00% | 0.00 | 00:00:00 | 0.00% | 0 (0.00%) |</p>
<table>
<thead>
<tr>
<th>Source/medium</th>
<th>Page</th>
<th>Acquisition</th>
<th>Behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>/bitstream/handle/2384/294701/Contextualised approaches to widening participation a comparative case study of two UK universities.pdf</td>
<td>1. scholar.google.co.uk</td>
<td>Users: 82 (0.95%) New Users: 82 (0.41%) Sessions: 80 (0.98%) Bounce: 0.00% Avg for view: 0.74% Pages/Session: 0.00 Avg Session Duration: 0:00:00</td>
<td></td>
</tr>
<tr>
<td>/bitstream/handle/2384/294812/Disrupting the 'whiteness' of field work in geography.pdf</td>
<td>2. scholar.google.com</td>
<td>Users: 5 (6.10%) New Users: 5 (6.10%) Sessions: 5 (6.10%) Bounce: 0.00% Avg for view: 0.17% Pages/Session: 0.00 Avg Session Duration: &lt;0:00:00</td>
<td></td>
</tr>
<tr>
<td>/bitstream/handle/2384/294757/Cross-modal face identity aftereffects and their relation to priming.pdf</td>
<td>3. scholar.google.co.uk</td>
<td>Users: 4 (4.88%) New Users: 4 (4.88%) Sessions: 4 (4.88%) Bounce: 0.00% Avg for view: 0.16% Pages/Session: 0.00 Avg Session Duration: 0:00:00</td>
<td></td>
</tr>
<tr>
<td>/bitstream/handle/2384/294812/Disrupting the 'whiteness' of field work in geography.pdf</td>
<td>4. scholar.google.co.uk</td>
<td>Users: 4 (4.88%) New Users: 4 (4.88%) Sessions: 4 (4.88%) Bounce: 0.00% Avg for view: 0.17% Pages/Session: 0.00 Avg Session Duration: 0:00:00</td>
<td></td>
</tr>
<tr>
<td>/bitstream/handle/2384/295314/re-economic-impacts-climate-change-agriculture-russia-010413-en+(1).pdf</td>
<td>5. scholar.google.com</td>
<td>Users: 4 (4.88%) New Users: 4 (4.88%) Sessions: 4 (4.88%) Bounce: 0.00% Avg for view: 0.17% Pages/Session: 0.00 Avg Session Duration: 0:00:00</td>
<td></td>
</tr>
<tr>
<td>/bitstream/handle/2384/294768/Crowding in children's VA tests.pdf</td>
<td>6. scholar.google.com.au</td>
<td>Users: 4 (4.88%) New Users: 4 (4.88%) Sessions: 4 (4.88%) Bounce: 0.00% Avg for view: 0.17% Pages/Session: 0.00 Avg Session Duration: 0:00:00</td>
<td></td>
</tr>
</tbody>
</table>
Email Report: Referral Traffic

From: bluyten@gmail.com
To: 
Subject: Google Analytics: Referral Traffic
Attachments: Google Analytics: Referral Traffic
Frequency: Once

Ik ben geen robot

Send  Cancel
Use?

1. Get a feel for what your "normal" volume of Google Scholar traffic is on a day
2. Setup a recurring report so that you can easily CHECK if suddenly you get diversions from this standard amount

IF you care about Google Scholar as a source of traffic, this is a way to stay on top of the flow of incoming traffic, and to start looking for sources of problems when you see alarming drops in traffic.
Questions?

If you want these slides, please take 5 minutes to give us feedback on this webinar. The link to the slides is shown after completing the questions.

bit.ly/atmire-scholar-webinar-feedback
Image credits

Web Crawler - Scot Swigart https://flic.kr/p/aoYNg

Mutianyu Great Wall - Arian Zwegers https://flic.kr/p/atS39L