Good morning, good afternoon and good evening, as we have a full house of over 100 Atmire client attendees from over the globe. It's the very first time that we are doing a webinar like this, and we are over the moon that so many of you responded with great interest to this initiative.

We WILL attempt to record the webinar, but it's only until that is completely successful that we will be sure that we'll be able to circle out the recording. If you want, you can definitely get the slides and the speaker notes.

For the people on the call who don't know me and are unfamiliar with my obsession with indexing and google scholar in particular: I have been interested in and working on this topic for many years. This lead to participating in the Open Repositories developer challenge back in 2014, where I demonstrated an initial prototype of the Analyzer software that I will be showing you later.

Please be aware that I do not work for Google, and that in many ways Google Scholar is still a black box for me as well. However, I've had a very good relation with people in the Google Scholar team over the years, who were kind enough to review these slides and didn't find any major mistakes in what I'm about to present to you.
But first

Does Google Scholar rely on OAI-PMH?

While many crawlers, including OpenAIRE, BASE and CORE rely on OAI-PMH, Google scholar does not. So your OAI-PMH configuration or related challenges can have no impact on your google scholar indexing.
What does the "site:" operator tell you about coverage of your repository?


Keep in mind that the result count of the "site:" operator is not a good indicator of coverage of your website in Google Scholar. First, this operator currently only searches primary versions of the papers. If you're not the primary publisher, some of the papers that you host may not be counted. Second, the result count is usually estimated based on searching a small fraction of the index (the purpose of the result count is to help users refine their queries and not coverage checking). As a result, this estimate may not be accurate. If you're alarmed that the result count for your site is low, please confirm the problem with a more detailed check. We recommend trying to find several dozen of sample papers using search by title.

Does Google Scholar index all your items?

No, as we will see later, Google Scholar is only interested in, and will only index scholarly content. Also, files that are bigger than 5MB are better uploaded to Google Books.

How long does it take for an item to get indexed?


Keep in mind that changes that you make on your website will usually not be reflected in Google Scholar search results for some time. New papers are normally added several times a week; however, updates of papers that are already included usually take 6-9 months. Updates of papers on very large websites may take several years, because to update a site, we need to recrawl it - the time it takes to recrawl a large site is usually limited by the speed at which the target website is able to deliver content to the search robots.
If you don't see your item in the search results, does that mean it isn't indexed?

As we will see later, make sure you check the "all versions" link. At first, Google Scholar shows the primary version of an item. Other versions are available through that link.

The only way to make a sure statement about an item NOT being indexed, is if you search for a URL to an item page and it gives you 0 results. As soon as you get a result, it means that Google Scholar is aware and has indexed your item URL.

Is it required to register your repository somewhere in order to get it included?

The answer is NO, such a registration is NOT required. HOWEVER, there IS a registration form where you CAN register your site. Normally, if you adhere to the following site wide basic principles, there is no need to register your site.

But if your repository is brand new, it doesn't hurt to fill out the form anyway. This also has the added benefit that Google will have you as a contact person for the repository.


If you've met all content, crawl, and indexing guidelines, but your papers still aren't indexed correctly after six months, then it's possible that the search robots aren't aware of your website. To rectify this, you can request manual configuration of your website. Content reviews are normally completed within six weeks from the time you submit your request.

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"
BAM, if you adhere to these site wide principles, you actually open up the gates and give Google Scholar a shot at finding and indexing your website.

The only right test to verify if an item is indexed

1. you take the url to an item page
2. you search for that particular url, in google scholar

If you get something back, Scholar knows about the URL
If you are not getting something back, the URL is not indexed.

It's that simple.

There is no known method to verify the coverage, in batch, for ALL of your repository contents. So the method you need to apply is that you sample as many items you deem necessary.
Finding items that are NOT being indexed, and especially establishing a PATTERN between those items, is the key to finding the root cause and optimizing your indexing.

Here you see this at work: I have searched for the URL and by retrieving a result, it is a positive confirmation that the URL is indexed.

You notice that the result & the file link shown is NOT the repository result. This is because Google Scholar doesn't want to show you duplicates for the same item.
It tries to find one occurrence of the item that it regards as the primary one, and then clusters all of the other versions together.
These versions are ALSO available to you if you click "All X versions"
In these clustered results, you do see the repository record, but it doesn't have a file next to it. This brings me to the point of what content is exactly indexed by Google Scholar.

**Item specific principles**


As you can see here, datasets are not being picked up by Google Scholar. There is a now a dedicated dataset search from Google. Getting your repository contents.
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

This is the explanation why my test item had no link next to it, because the repository version only had XML uploaded with it, not HTML or PDF.

The next slide is the most important slide in this webinar

Metadata Matters!

Required fields
- citation_title
- citation_author
- citation_publication_date
- citation_pdf_url

A KEY knockout criterium is AUTHOR ORDER: Google Scholar expects that the order of the authors that you have in the metadata, matches the order of the authors that appear in those first pages of the full text.

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.

---

Special metadata tags expressed in <HEAD> of the page

A KEY knockout criterium is AUTHOR ORDER: Google Scholar expects that the order of the authors that you have in the metadata, matches the order of the authors that appear in those first pages of the full text.

---

Metadata Example

```xml
<meta name="citation_title" content="The tertiary structure of the phosphoglycerate kinase regulator domain (KDP1) place a critical role in regulation of Epyromon mobilization in dormant leaf">
<meta name="citation_author" content="Liu, X.">
<meta name="citation_author" content="Liu, X.">
<meta name="citation_publsh" content="Palaeontology, Mary">
<meta name="citation_author" content="M. Blue, Name A.">
<meta name="citation_author" content="M. Blue, Name B.">
<meta name="citation_author" content="M. Blue, Name C.">
<meta name="citation_publication_date" content="2018/05/17">
<meta name="citation_journal_title" content="Journal of Biological Chemistry">
<meta name="citation_author" content="1">
<meta name="citation_language" content="en">
<meta name="citation_pdf_url" content="https://example.com/retrieve/2020/11/16/full.pdf">
```

---

Metadata mapping configuration

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

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

google.citation_pdf_url = dc.identifier

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

---
Journal and Conference Papers

Optional fields
- citation_journal_title, citation_conference_title
- 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


Basic Principles - Item specific

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

When you enter the site you can see that more than 2400 repositories have been checked, and that in total, 30 million items are hosted on these repositories.
The buttons on top allow you to jump to different sections of the page.

Anyone can use the tool for analysing any repository. You can either enter a URL in the Get started section. Or click on one of the recently analysed repositories.
The initial scan that the tool makes, is whether the repository responds fast enough and if the site can be recognized as a DSpace repository. Clicking on the "Start Extended Analysis" button on the top right, allows you to fill out your email address to get a more detailed analysis.

If the site-wide principles are followed and respected, all of the extended analysis boxes at the bottom of the page should be green. Only then, the functionality is unlocked to analyse individual items.

As indicated earlier, testing specific items will have little chance of success, if there are site wide problems with your repository.

I have the test item we used earlier in the presentation, a valid DSpace item, but with an XML file instead of a PDF full text file. Let's have a look at how the scholar metadata looks in the HEAD section on the source of the page.
After the item has been analysed, you can see the title and the authors. If a link was found to the pdf, it's shown there, as well as the link to the item page itself.

If the automated analysis runs well, the tool can determine by itself if the item and/or the full text shows up in Google Scholar. However, due to Captcha's and other measures against robots, the automated analysis often doesn't find the item or the full text anymore. The next few tools allow you to test for yourself, and to indicate whether item metadata or full text has been found.
In this box that tells you that the tool hasn’t found the item metadata yet, there are two links: one to execute the query in your browser, searching the repository item URL in Google Scholar. With the second link, YOU tell the tool that you have found the item metadata.

Let’s look at a repository for which a number of items have already been analysed.

All of the items that have been analyzed, are summarized on the page of the repository. When the title is red, the metadata hasn’t been found in Google Scholar. When the file icon is red, the file hasn’t been found in Google Scholar.
Page views and downloads can be accompanied with a "referral" field that identifies from which website a user is coming. This field is not always filled out, so any data you see here are lower estimates/underestimates of traffic actually getting from a source.

Now that we have narrowed down to a selection of traffic coming from Google Scholar, we can look at:
- which items were downloaded
- which (academic networks) are the users coming from
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.

Other than satisfying endless streams of random curiosity (GUILTY AS CHARGED),
Before opening up the chat for questions, just one anecdote to emphasize that everything I've told you today is very general. There can be many, very specific and unique challenges that prevent your items from being indexed.

A couple of weeks ago, when satisfying my endless curiosity, I was running the analyzer against the largest repository I know, the Chinese Academy of sciences IR Grid. This DSpace has over a million items.

None of them show up in Google Scholar, and I think I figured out why.

A normal HTTP request fetches the entire page, and also gives you an HTTP status code. If you are interested in very quickly getting the status code, but don't want to put the load of serving a full page (in your own crawler) or on the site you are hitting, you can issue a request that ONLY serves the headers, a so called HEAD request.

The IR Grid’s tomcat is configured in such a way, that it denies these head requests. Result: even if they would comply with all of the other guidelines I mentioned so far, not a single item would get indexed.