Google Scholar
Vs Institutional Repository

Tue Aug 27 2019, 2:30pm–3:30pm, Max Atria Garnet 219

Veerle kerstens
veerle.kerstens@kuleuven.be

Mehmet Celik
mehmet.celik@kuleuven.be
@celik.be
https://github.com/mehmetc
Contents

- Introduction
- Issue: making our institutional repository discoverable in Google Scholar
- What we did
- Conclusion: experiences
- Possible plans for the future
LIBIS is a department of KU Leuven and supports central ICT applications for the management & discovery of information sources for libraries, archives, museums and other (heritage) collections (GLAM).

Division of Leuven Research and Development (LRD)

We provide consultancy on all Ex Libris products.
Institutional Repositories

The Stakeholders

- Researchers:
  If publications are better findable -> more reads
  -> more citations -> more recognition -> more funding

- Institution:
  ○ More citations gives a better ranking
  ○ More visibility and credibility

- Collection managers: special collections
Lirias
Institutional bibliography and repository of KU Leuven Association

- > 550,000 publications
- > 120,000 with files uploaded
- > 60,000 in Open Access
- Metadata managed in Elements
- Full text stored in DSpace
- Primo as end user interface
DokS
Repository of bachelor theses of colleges in Flanders

• > 25.000 bachelor theses
• > 20.000 in Open Access
• Metadata managed in Alma
• PDF’s stored in Rosetta
• Primo as end user interface
Publication Life Cycle
Google and Google Scholar: little in common…

Crawling

Indexing

Ranking algorithms
Crawling

Google Scholar

- Scope: scholarly publications (no books)
- No big documents (?)
- Crawls documents and bibliographic information
- Format: pdf or html
- Full text or at least abstracts needed (?)
- Updates records: 6-9 months to a year.

Google

- Scope: “all”
- Crawls documents
- Many formats
- Updates records dependant on the importance of the site
- Follows any hyperlink
Indexing

- Meta-tags: format Highwire Press, Eprints, BE Press or PRISM
- Grouping of results

- Meta-tags: format schema.org
- Every url displayed separately in search results => issue of duplicates
- Version of the publisher is the primary version.
- Also pre-prints are grouped with the final version.
Other conditions

- Requirement: PDF in the same subdirectory/domain as metadata
- 1 PDF per document
- PDF’s are text searchable (no scans)
- At least abstract is available (?)
- No pop-ups or requirements to log-in
- ...

Indexing Guidelines

Google Scholar uses automated software, known as “parsers”, to identify bibliographic papers, as well as references between the papers. Incorrect identification of bibliographic data would lead to poor indexing of your site. Some documents may not be included with incorrect author names or titles, and some may rank low because the bibliographic data would not match (correct) references in the papers. To avoid such problems, you need to provide bibliographic data and not “automated” “parser” software can process.

1. Preparing article URLs

Place each article and each abstract in a separate HTML or PDF file. At this time, effectively index multiple abstracts on the same webpage or multiple papers in the same directory, we’re unable to index different sections of the same paper in different files. Each document must have its own unique URL in order for it to be included in Google Scholar.

2. Configuring the meta-tags

If you’re using repository or journal management software, such as Eprints, DSpace, or OJS, please configure it to export bibliographic data in HTML ("meta") tags. The database supports HigherWire Press tags (e.g., citation_title), Eprints tags (e.g., eprints_title), and PRISMD tags (e.g., prism_title). Use Dublin Core tags as a last resort - they work poorly for journal papers because Dublin Core doesn’t include all the details, such as journal title, volume, issue, and page numbers. To check that these tags are present, you can view the HTML source.

A. The title tag, e.g., citation_title or DC title, must contain the title of the paper, book, or journal in which the paper was published, or for the repository. This tag is required for inclusion in Google Scholar.

B. The author tag, e.g., citation_author or DC Creator must contain the actual authors of the paper. Don’t use it for the author of the website or the first author, e.g., thesis advisors. Author names can be listed either as "John Smith" or "Smith, John". Each author name must be in a separate tag and omit all affiliations and addresses. At least one author tag is required for inclusion in Google Scholar.

C. The publication date tag, e.g., citation_publication_date or DC issued, must be included, i.e., the date that would normally be cited in references to the papers. Don’t use it for the date of entry into the repository - that should go in the citation_online_date instead. Include full dates in the “2010/5/12” format, for example. This tag is required for inclusion in Google Scholar.

D. For journal and conference papers, provide the remaining bibliographic details.
Crawling & Indexing: monitoring

- No systematic checks possible
- Only via samples
- Contact person at Google Scholar

- Google Search Console to monitor process of crawling & indexing
Relevance ranking

- Number of citations +++ (1)
- Search term in title ++
- Author name, Journal title++
- Search terms in document +
- Reputation of author, Journal+
- Date +- 
- Number of times the search term appears in the document –
- Synonyms in document -

(1) Rovira, Cristòfol; Guerrero-Solé, Frederic; Codina, Lluís (2018). “Received citations as a main SEO factor of Google Scholar results ranking”. El profesional de la información, v. 27, n. 3, pp. 559-569. https://doi.org/10.3145/epi.2018.may.09
How to make our publications discoverable in Google Scholar?

Facilitate crawling of Primo?

- Required meta tags currently not available. (schema.org tags are not supported)
- Requirement of Google Scholar to have full text and metadata in the same domain

=> We developed our own solution
The institutional repository

The players

METADATA

DOCUMENTS

ENDUSER UI
Crawlers visiting

sitemap.xml

Resolv
Biomechanical changes in articulation of the jaw joint due to aging

Publication date: 2018-10

Biomechanical changes in articulation of the jaw joint due to aging

Author:
Mirahmadi, F

Abstract:
Age-related thinning of articular cartilage is supposed to increase susceptibility for osteoarthritic changes in the temporomandibular joint. The present project will test the hypothesis that this process is caused by an increased stiffening of the cartilage. This stiffening is considered to result in a reduction of the possibilities for metabolic activity of the chondrocytes in the layer near the subchondral bone, followed by a subsequent calcification of this area. This hypothesis will be tested (i) by experimental analysis of propagation of solutes in the cartilaginous structures of the jaw joint during aging, and (ii) by quantification of the mechanical changes in aging cartilage. (iii) Consequently, experimental testing of the artificial aging of cartilage will be performed to examine the consequences of crosslinking increase in aging cartilage.
Biomechanical changes in articulation of the jaw joint due to aging

Publication date: 2018-10

Author:
Mirahmadi, F.

Abstract:
Age-related thinning of articular cartilage is supposed to increase susceptibility for osteoarthritic changes in the temporomandibular joint. The present project will test the hypothesis that this process is caused by an increased stiffening of the cartilage. This stiffening is considered to result in a reduction of the possibilities for metabolic activity of the chondrocytes in the layer near the subchondral bone, followed by a subsequent calcification of this area. This hypothesis will be tested (i) by experimental analysis of propagation of solutes in the cartilaginous structures of the jaw joint during aging, and (ii) by quantification of the mechanical changes in aging cartilage. Consequently, experimental testing of the artificial aging of cartilage will be performed to examine the consequences of crosslinking increase in aging cartilage.
Biomechanical changes in articulation of the jaw joint due to aging

Age-related thinning of articular cartilage is supposed to increase susceptibility for osteoarthritic changes in the temporomandibular joint. The present project will test the hypothesis that this process is caused by an increased stiffening of the cartilage. This stiffening is considered to result in a reduction of the possibilities for metabolic activity of the chondrocytes in the layer near the subchondral bone, followed by a subsequent calcification of this area. This hypothesis will be tested (i) by experimental analysis of propagation of solutes in the cartilaginous structures of the jaw joint during aging, and (ii) by quantification of the mechanical changes in aging cartilage. (iii) Consequently, experimental testing of the artificial aging of cartilage will be performed to examine the consequences of crosslinking increase in aging cartilage.
Bitstream proxy
Statistics
GoAccess on HTTP server logs for simple reports

OVERALL ANALYZED REQUESTS

- Total Requests: 205,935
- Valid Requests: 205,922
- Failed Requests: 13
- Init. Proc. Time: 0 secs
- Unique Visitors: 21,605
- Requested Files: 66,880
- Excl. IP Hits: 0
- Referrers: 0
- Not Found: 1,034
- Static Files: 0
- Log Size: 0 Byte
- Tx. Amount: 0 Byte

18/JUL/2019 – 21/AUG/2019

UNIQUE VISITORS PER DAY - INCLUDING SPIDERS

Hits having the same IP, date and agent are a unique visit.

# Hits Visitors Data

205,922 21,605 35 Total
Max: 34,830 Max: 851
Min: 298 Min: 33
Statistics
Kibana and Resolver logs for more in depth analysis
How to deal with records that have more than one file attached?

Google Scholar can only link to one file

- Freely available and restricted version?
  - => only freely available version is exposed for crawling

- More than 1 freely available file?
  scenario's:
  - Files are different parts of the work
  - Files are different versions of the work
Files are different parts of the work

Advise to researchers not to split up their work into parts.
Files are different versions of the work

Currently no possibility to indicate in the metadata which file is the preferred one

- if accepted and published version are available
  => published version is selected for crawling

- Otherwise: all files are presented for crawling
  => first occurrence is taken by Google Scholar
What to do with dead links

Can happen if

- Failed to set embargo
- Metadata gets replaced

Possible consequences

- Lower ranking of repository data
- Metadata removal
- Link roth

Send 410: “Gone” instead or better forward to the correct entry
Our experiences

- SEO for Google Scholar appeared to be a long lasting process
- What helped us to solve issues:
  - Systematic checks in GS
  - Monitoring of the harvesting
  - Contact person at Google Scholar
- Advice on best practices for researchers is important
Best practices for researchers

- Good metadata is important, always provide an abstract.
- Quality of the uploaded files is important.
  - Don’t upload image-scans of papers
  - Need of well-formed citations section in the document
- Do not upload more than one freely available file per paper
- Open Access papers will get more exposure

Record with an abstract

Understanding social innovation as an innovation process: Applying the innovation journey model
PRA Oeij, W van der Torre, F Vaas... - Journal of Business ..., 2019 - liras.kuleuven.be
The innovation journey is a process model distinguishing between the initiation, developmental and implementation/termination period of innovations; it looks at drivers and barriers, like innovation managers, investors, setbacks, adaptation, infrastructure. We ...

Record without an abstract

Views and attitudes towards Evidence-Based Practice in a Dutch social work organization
R Van der Zwt, D Beneken genaamd Kolder... - Journal of Evidence ... - liras.kuleuven.be
Journal of Evidence-Based Social Work. ISSN 1543-3714. Publisher: Taylor & Francis (Routledge). Views and attitudes towards Evidence-Based Practice in a Dutch social work organization. Author: Van der Zwt, R. Beneken genaamd Kolder, D. Schalk, R.; Van ...
Possible plans for the future

- Implement the landing pages for our repository of Master Theses DokS
- Use the monitoring tool to report on the use of repository items (page views and full text downloads)
- Facilitate crawling of by Google (include schema.org tags in the landing pages)
- Landing page as default?
Thank you

Questions ?