* Metadata for Web Crawler

Volker RW Schaa
GSI, Darmstadt, Germany

Team Meeting 2026 — KEK, Tsukuba, Japan

* Metadata

* In July 2006 I started adding metadata in the scripts to get better search results from Google (specifically Google Scholar)
  * first trial: I searched for metadata in web pages by Springer, Elsevier, etc.
  * citation_...: found citation_ tags and included them

* Over the years this became clearer
  * Dublin Core: found DC. keywords and included them to
  * Highwire Press Tags came in March 2016 after I found these explicitly mentioned as critical
  * Tags: 22 citation_ tags + the corresponding DC.-tags

* Metadata

* Caveats
  * citation_pdf_url: Google Scholar will not index the entry,
    if this link does not point to the full-text PDF
  * URLs: html for landing pages (DOI) and metadata must be permanent and static.
  * robots.txt: check for free access to all pages carrying <metadata> or
    provide direct links to the material
* Links
  * links: most links I used at that time are not accessible anymore
    OK ⇒ https://scholar.google.com/intl/en/scholar/inclusion.html
  * survey: https://www.monperrus.net/martin/accurate+bibliographic+metadata+and+google+scholar
* AI
  * Today when you ask ChatGPT or any other AI you get all these information directly

*What happens when my conference finishes, and why?

Metadata

Volker RW Schaa
GSI, Darmstadt, Germany

Team Meeting 2022 — Knoxville, TN, USA
Thursday, 2022-11-10T11:00

* Metadata

* XMP module: Meta data incorporated in PDF files:
  * pdfauthor: list of authors (with Unicode alternative when available)
  * pdfcreationdate
  * pdfkeywords: list of found keywords
  * pdflang: English
  * pdfsubject: main classification [+ sub classification]
  * pdftitle: title of paper
  * pdfcontacturl: https://www.jacow.org
  * pdfcopyright: CC by 3.0/4.0 .....
  * pdfdoi: 10.18429/JACoW-<conference>-<paper code>
  * pdfeissn: electronic series ISSN
  * pdfisbn: conference ISBN
  * pdfisbn: conference ISBN
  * pdflicenseurl: https://creativecommons.org/licenses/by/3.0/ or /4.0/
  * pdfnumpages: start page + end page in proceedings of conference

Metadata (cont.)

* pdfnumpages: start page + end page in proceedings of conference
* pdfpublication: Proceedings of <conference_name>
* pdfpublisher: JACoW Publishing, Geneva, Switzerland
* pdfpubtype: Journal
* pdfurl: `https://www.jacow.org/<conference_name>/papers/<paper code>.pdf` points to the complete URL of the document
* pdfpubstatus: not filled (yet) as it provide “states” of a PDF which might conflict with our final version concept, as JACoW is not a pre-press publication cite like arXiv. Possible values:
  * »AO« Author’s Original
  * »SMUR« Submitted Manuscript Under Review
  * »AM« Accepted Manuscript (best fitting type)
  * »P« Proof
  * »VoR« Version of Record
  * »CVoR« Corrected Version of Record (might sometimes apply after correction)
  * »EVoR« Enhanced Version of Record

Metadata (cont.)

* Highwire Press Metatag: Metadata incorporated in web pages [all pages with bibliographic data for one submission] for search-engine and AI driven discovery (i.e. Google Scholar, PubMed, and other scientific web crawlers).
* see for example view-source:https://jacow.org/ipac2022/export/THPOPT014-bib.htm
* <meta name="citation_title" content="Simulation and Optimization of SPS-II Linac" />
* <meta name="citation_author" content="Chanwattana, Thakonwat" />
* <meta name="citation_author" content="Chunjarean, Somjai" />
  * more co-authors
* <meta name="citation_date" content="2022/07" />
* <meta name="citation_publication_date" content="2022/07/01" />
* <meta name="citation_publisher" content="JACOW Publishing, Geneva, Switzerland" />
* <meta name="citation_firstpage" content="2590" />
* <meta name="citation_lastpage" content="2593" />
* <meta name="citation_doi" content="10.18429/JACoW-IPAC2022-THPOPT014" />
* <meta name="citation_isbn" content="978-3-95450-227-1" />
* <meta name="citation_issn" content="2673-5490" />
* <meta name="citation_keywords" content="linac, simulation, synchrotron, emittance, storage-ring" />
* <meta name="citation_language" content="en" />

* Metadata (cont.)

* <meta name="citation_conference_title" content="13th Int. Part. Accel. Conf. (IPAC'22), Bangkok, Thailand, 12-17 June 2022" />
* <meta name="citation_pdf_url" content="https://jacow.org/ipac2022/papers/thpopt014.pdf" />

These are the standard tags of Highwire Press. In addition some other tags are supplied as recommended by Google Scholar: but these are just repetions of the previously given metadata tags under a different name schema (DC: Dublin Core).

```
<meta name="DC.creator" content="Chanwattana, Thakonwat" />
    more "DC.creator" records
<meta name="DC.issued" content="2022/07" />
<meta name="dcterms:issued" content="2022" />
<meta name="DC.title" content="Simulation and Optimization of SPS-II Linac" />
<meta name="DC.publisher" content="JACOW Publishing, Geneva, Switzerland" />
<meta name="DC.citation.spage" content="2590" />
<meta name="DC.citation.epage" content="2593" />
<meta name="DC:identifier" content="info:doi/10.18429/JACoW-IPAC2022-THPOPT014" />
<meta name="dcterms:isPartOf" content="urn:ISSN:2673-5490" />
<meta name="dcterms:isPartOf" content="urn:ISBN:978-3-95450-227-1" />
<meta name="DC.subject" content="linac, simulation, synchrotron, emittance, storage-ring" />
<meta name="citation_language" content="en" />
<meta name="dcterms:bibliographicCitation" content="13th International Particle Accelerator Conference (IPAC'22), Bangkok, Thailand, 12-17 June 2022" />
<meta name="DC.identifier" content="https://jacow.org/ipac2022/papers/thpopt014.pdf" />
```

* Acronyms

DC

* Dublin Core
  also known as the **Dublin Core Metadata Element Set** (**DCMES**), is a set of fifteen “core” elements (properties) for describing resources. This fifteen-element Dublin Core has been formally standardized as ISO 15836