Towards a Comprehensive and Up-To-Date Institutional Repository

DEVELOPMENT OF A PUBLICATIONS TRACKING PROCESS

Mohamed Baessa, Daryl Grenz, Han Wang

June 15, 2016
Comprehensive? Up-To-Date?

- A complete and reliable resource for research produced by KAUST researchers or supported by KAUST funding.
Why?

- June 2014 adoption of institutional open access policy
- Library task: “develop and monitor a plan to comply with this policy and existing copyright obligations in a manner as convenient for the faculty as possible.”
Is our repository ready?

- Waiver form option
- Type-based submission forms
- Cover sheet with version, rights and citation
- Faceted search
  - Distinguish downloadable from embargoed or metadata-only items.

**Item Availability:**
- Open Access (3386)
- Embargoed (839)
- Metadata Only (7481)
Plan?

• Evaluate systems for:
  – harvesting publication metadata
  – depositing full texts as appropriate

• Select one and integrate it with the existing repository.
Plan carried out…?

• Subsumed into larger project:
  – Selection and implementation of research information system (CRIS).

• Only one consideration among several
  – Institutional reporting, grant management …
In the meantime…

At least we should have a copy of everything with a license that permits redistribution.

- Pubmed Central open access subset retrieval and batch ingest.
  - 300+ publications in simple archive format
- Compared Scopus export with Sherpa Romeo and DOAJ to identify other articles with open access licenses
  - 150+ publications, deposited manually
- Set up Google Scholar email alert for “Creative Commons”
  - 3-5 publications a week
March 2015

How are we doing with the open access policy?

• Are people self-depositing?
  ➢ Same as before OA policy, primarily in electrical
    engineering and applied mathematics (“send us arXiv IDs”).

• One research center
  ...
    “help us make sure everything is in the repository”

-- We don’t know what we’re missing --
Contacting faculty individually

- List to each faculty member of publications in the last nine months showing what was deposited and what was not.
  - By Scopus ID
    - Then checking on repository by DOI
  - Manual highlighting and review
  - Email with explanation of open access policy and publisher policies
Responses to contacting faculty individually

• Your list is incomplete:
  – please search on Web of Science
  – look at my Google Scholar profile
  – see my attached CV.

• Why are you contacting me, you can download these from the publisher.

• I don’t think the publisher allows this.

• I don’t keep old versions of papers, ask my co-authors.

• No response…
Responses to contacting faculty individually

• My accepted manuscripts are all on my web site / arXiv, please copy from there.
• Requested files attached.
• I always pay to publish OA, please check individual article licenses, not just journal policies.
• 650 articles requested from 130 faculty, 60+ items deposited.
How can we do better?

Maximize what we do on researchers’ behalf:
• Deposit when possible.
• Check compliance with publisher policies and embargoes.

Contact authors closer to the time of publication:
• All KAUST authors at once.
Commit to being a complete resource for KAUST work:

• Add metadata-only records to the repository.

• Emphasize that the information will be reused:
  - ORCID
  - PlumX

• All item types?
  - Open access policy applies only to articles, conference papers and book chapters.
  - But repository will also track patents and meeting abstracts
What source(s)?

- Exports from Scopus and Web of Science
  - Varied handling of affiliations
    - Scopus treats listing of current or present address as normal affiliation
    - Web of Science ignores current and present addresses
  - How successfully are affiliation variants matched to institutions?

- "Export" from Google Scholar
  - Full-text search…
  - Using Publish or Perish software
  - Broken into manageable batches
What about outputs from KAUST funded research at external institutions?

- Same sources, with the addition of Fundref/Crossref

Essentially the same conclusion, no source is comprehensive, reliable and up-to-date enough for us to feel confident in relying on it alone.
Established tracking process:

- **PHP scripts daily check:**
  - **Indexer APIs:**
    - Scopus, Web of Science, Google Scholar (using scholar.py)
  - **Publisher APIs:**
    - IEEE, Sciencedirect, Springer, Nature
  - By affiliation keywords and name variants
How does that work:

• De-duplication based on DOI?
  – Spot the difference?

<table>
<thead>
<tr>
<th>DOI</th>
<th>DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1016/S0065-2881(10)57009-3</td>
<td>10.1016/B978-0-12-381308-4.00009-1</td>
</tr>
<tr>
<td>10.1364/OE.23.0010224</td>
<td>10.1364/OE.23.010224</td>
</tr>
<tr>
<td>10.1021/acs.organomet.5600749</td>
<td>10.1021/acs.organomet.5b00749</td>
</tr>
</tbody>
</table>

• Both Scopus and Web of Science have small numbers of invalid DOIs.
How does that work:

• De-duplication based on DOI?
  – Cases of multiple DOIs:
    • *Angewandte Chemie*
      – German and international editions issue separate DOIs, though research article is in English in both cases.
    • **Cover pages**
      – Issued their own DOIs, and citations sometimes use this DOI, rather than the main article DOI.
    • **IEEE**
      – Initial error of year in initial DOI registration for all items in a conference.
• Items without DOIs?
  – From WOS and Scopus
    • Mostly meeting abstracts, but also a small number of items for which we can identify valid DOIs.
  – From Google Scholar
    • Attempted matching to DOIs based on title and known publisher URL strings
  – Self-deposited accepted manuscripts prior to DOI issuance?
How does that work:

• Centrally managed deposit
  – Library staff check that there is a KAUST affiliation or KAUST funding acknowledgement
  – Deposit if license or publisher policy permits
    • Including accepted manuscripts made available by publishers prior to copyediting and final formatting.
    • With embargoes as required
How does that work:

• Manuscript requests
  – If only publisher version is available and its deposit is not permitted.
    • Request accepted manuscript from KAUST authors
    • If file sent, deposit.
    • Else, upload metadata-only record via batch CSV upload after two weeks.
## Time lag of Scopus (for sample of recent items)

<table>
<thead>
<tr>
<th>First source</th>
<th>Average lag (months)</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>2.5</td>
<td>35%</td>
</tr>
<tr>
<td>Scopus</td>
<td>0</td>
<td>27%</td>
</tr>
<tr>
<td>Springer</td>
<td>2</td>
<td>6%</td>
</tr>
<tr>
<td>IEEE</td>
<td>4</td>
<td>15%</td>
</tr>
<tr>
<td>Sciencedirect</td>
<td>1.7</td>
<td>13%</td>
</tr>
<tr>
<td>Nature</td>
<td>1.5</td>
<td>3%</td>
</tr>
</tbody>
</table>
### Overall compliance with the open access policy

<table>
<thead>
<tr>
<th></th>
<th>Number of Publications</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>After July 2014</td>
<td>2436</td>
<td>100%</td>
</tr>
<tr>
<td>File deposited</td>
<td>1804</td>
<td>74%</td>
</tr>
<tr>
<td>No file deposited</td>
<td>632</td>
<td>26%</td>
</tr>
</tbody>
</table>
Versions and licenses:

## Versions and licenses of deposited files

<table>
<thead>
<tr>
<th>Version Type</th>
<th>Number of Publications</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Files deposited</td>
<td>1804</td>
<td>100%</td>
</tr>
<tr>
<td>Publisher version with CC license</td>
<td>502</td>
<td>28%</td>
</tr>
<tr>
<td>Publisher version deposit allowed by journal policy</td>
<td>327</td>
<td>18%</td>
</tr>
<tr>
<td>Accepted manuscript deposited</td>
<td>953</td>
<td>53%</td>
</tr>
<tr>
<td>Other versions or licenses</td>
<td>22</td>
<td>1%</td>
</tr>
</tbody>
</table>
Effectiveness of email requests:

Response to emails to all KAUST authors

<table>
<thead>
<tr>
<th>Description</th>
<th>Number of Publications</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items for which email manuscript request was sent to all KAUST authors</td>
<td>904</td>
<td>100%</td>
</tr>
<tr>
<td>Manuscripts deposited after receipt from one of the authors</td>
<td>326</td>
<td>36%</td>
</tr>
<tr>
<td>No manuscript from authors after request</td>
<td>578</td>
<td>64%</td>
</tr>
</tbody>
</table>
Where will it all end:

• More item types?
  – Patents
  – Meeting abstracts / presentations
  – Datasets
  – Software

• More automation?
  – SWORD or Rest API deposits and metadata updates to DSpace?
Questions?

THANK YOU!