Machine accessibility of Open Access scientific publications from publisher systems via ResourceSync

Petr Knoth
Knowledge Media institute, The Open University
United Kingdom
Research literature contains some of the most important information we have assembled as human species, such as cures to diseases and answers to many of the world’s challenges we are facing today.
Reading and systematically analysing this information is beyond human capacities.
Why machine accessibility of publications?

- TDM can only fulfil its potential if TDM tools can be applied on the:
  - widest possible set of publications
  - as soon as publications are made available
- Many publication providers => need for interoperability
Expertise Directory 1/2

- Contacted publishers to clarify the expected approaches
- Developed code to implement them:
  - In most cases the final approach was different from the suggestions we received.
  - Tested the approach for scalability
- Documented the approach, justified why we followed it (including what did not work) and gave recommendations to publishers.

The expertise directory is available at: https://github.com/openminted/omtd-publisher-connector-harvester/blob/master/interoperability-layer/interoperability-layer.adoc
Example of limitations as described in the Expertise Directory for Elsevier:

<table>
<thead>
<tr>
<th>Task</th>
<th>Limitation</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieving list of all items</td>
<td>To get a list of all items available in Elsevier, one must crawl the sitemaps - impossible by querying the API</td>
<td>Build a custom crawler to get all items in Elsevier.</td>
</tr>
<tr>
<td>Updating a list of DOIs</td>
<td>Once the sitemaps have been harvested, it is not easy to get a list of additions since the last harvest without re-crawling</td>
<td>Though far from optimal solution, reharvesting sitemaps and comparing with old copy to discover new content, by the difference of both, is the solution followed by the current implementation.</td>
</tr>
<tr>
<td>Retrieval throttling</td>
<td>Elsevier API is has probably a throttling limit for the number of requests. We believe the limit is somewhere around 3 per second.</td>
<td>Unrestricting API usage; do not impose throttling limits.</td>
</tr>
</tbody>
</table>
The idea of the Publisher Connector

Provide seamless access over non-standard APIs

Why not OAI-PMH?

• Slow and very inefficient for big repositories.
• Standardised for metadata transfer but not for content transfer.
• Very difficult to represent the richness of metadata from a broad range of data providers.
The idea of the Publisher Connector

Provide seamless access over non-standard APIs

Why ResourceSync?
• Scalable implementation
• Metadata format agnostic
• Made for the Web
Integration with OpenMinTeD

Via CORE and the OMTD-SHARE schema.
How does it work?
Architecture 1/3

- Microservices architecture with a message queue as a communication channel.
- Discover, Retrieve, Expose (DRE) Workflow
Architecture 2/3

• Ingestion Services:
  • **Harvester service**: Discovers new resources (publications) and schedules them for downloading (via the message queue)
  • **Retriever service**: Retrieves scheduled publications from the queue and downloads them (both metadata and content) applying an appropriate data source download client for each publisher.
  • **Data source download clients**: publisher specific methods for discovery and retrieval + a generic CrossRef API wrapper.

• Exposure service:
  • **ResourceSync server service**: exposes publications according to the ResourceSync standard.
Architecture 3/3

- Message queue module: interface to a message broker (RabbitMQ) that is populated with publications events scheduled for downloading.
- Database module: Store and keeps downloads for incremental synchronisation.
Discovery

• Could be done via the CrossRef TDM API for some (typically smaller => scalability) publishers
  • Filtering by date of publication and for a set of OA licences
  • Sitemaps crawling for Elsevier

Diagram:
- Discovery
- Retrieval
- Expose
Retrieval

- Each publisher employs different methods and rules to download and retrieve an article.
## Scalability analysis

<table>
<thead>
<tr>
<th>Publisher</th>
<th>discovery</th>
<th>Metadata + Content (single thread)</th>
<th>Identification (is OA?)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elsevier</td>
<td>8m 1s</td>
<td>59m</td>
<td>instant</td>
</tr>
<tr>
<td>Springer</td>
<td>6m 52s</td>
<td>51m</td>
<td>n/a</td>
</tr>
<tr>
<td>Frontiers</td>
<td>16m 40s</td>
<td>2h 46m</td>
<td>n/a</td>
</tr>
</tbody>
</table>

* On a sample of 10k documents averaged over 2 trials

We can reprocess all Elsevier articles single threaded in about 100h
Exposure

- Scalable implementation of a ResourceSync server: For each publisher a new ResourceSync “capability” is created for its metadata and one for its content (pdf). The ResourceSync server is deployed at http://publisher-connector.core.ac.uk/resourcesync/
How many articles are provided as OA?

Discovery of OA articles (May 2017)

<table>
<thead>
<tr>
<th>Publisher</th>
<th>Metadata records</th>
<th>TDM-eligible license</th>
<th>OA articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Springer Nature</td>
<td>10,383,519</td>
<td>1,393,991</td>
<td>438,139</td>
</tr>
<tr>
<td>Elsevier</td>
<td>14,988,181</td>
<td>?</td>
<td>1,005,768</td>
</tr>
<tr>
<td>Frontiers</td>
<td>68,790</td>
<td>68,790</td>
<td>68,790</td>
</tr>
</tbody>
</table>

OA Percentage: 7% - TDM eligible 9.7%
## Volume sizes

As of August 2017

<table>
<thead>
<tr>
<th>Publisher</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elsevier</td>
<td>1,107,091</td>
</tr>
<tr>
<td>Springer</td>
<td>492,462</td>
</tr>
<tr>
<td>Frontiers</td>
<td>59,512</td>
</tr>
<tr>
<td>PLOS</td>
<td>172,812</td>
</tr>
</tbody>
</table>

Total: **1,831,877**
The OpenAIRE infrastructure has recently started a collaboration with the CORE Team led by Petr Knoth in order to include in the OpenAIRE metadata and file aggregation chain the resources made available via the ResourceSync connector realised at the CORE Team Lab. To this aim, the OpenAIRE team is testing the CORE ResourceSync Connector code with the intention of integrating it in its production system before the end of 2017.”

– Paolo Manghi – Technical Lead of OpenAIRE
Contributions

• Content:
  • We liberated over 1.8 million open access publications from publishers and made them available through a seamless layer.
  • As CORE integrates these papers, we have now over 8 million full-text papers in CORE.

• Technical:
  • First implementation and deployment of ResourceSync that scales to millions of items.
  • ResourceSync solves problems with aggregating content over OAI-PMH, faster & more efficient aggregation => fresher data in aggregators compared to OAI-PMH.

• More work in this direction upcoming as part of COAR NGR.