Systems Integration with Pure

Moving away from the Masterlist and other benefits of automation

Farah Karimi | Customer Consultant, North America
Oct 31, 2018
Agenda:

i. Advantages of aggregating data in Pure
ii. Ways to bring data into Pure
iii. Automation vs Data Quality
iv. Challenges with the Masterlist
Aggregate once, use many times for many things by many people

Pure

- Robust data model
- Customizable quality assurance workflows
- Interrelated Content

Outputs
- Showcasing
- Open Access to research
- Performance Dashboards
- Customized performance reports
- View research grant income success rates
- CVs
- ...

Who benefits
- Univ. Snr. officials
- Deans & dept heads
- Researchers
- Comms office
- Library
- Students
- Public & biz community
- Tech Transfer
- ...

Applications
- Grants
- Projects

Grants
- Research Output
- Activities
- Datasets

Persons
- Research Output
- Journals
- Prizes
- Course Taught

External Persons
- External Organizations

Organizational Units

Who benefits
Client use cases of data aggregation in Pure

I. a client is looking to **consolidate dispersed organizational hierarchy** and maintain it in a repository that would allow them to repurpose that data e.g.: on their department website

II. a client chooses to use Pure as a **publication repository**. Add data for electronic links, full text files, embargo dates, APC charges, license info, etc. within Pure’s research output template

III. At institutions that subscribe to both **SciVal and Pure**, the two systems can be integrated such that, SciVal can be used to understand global trends and Pure can be used for insights on operational efficiency

IV. **Metrics and Ids** coming in from integrations:

<table>
<thead>
<tr>
<th>Metrics/ ids</th>
<th>Source</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scopus author IDs, snip, sjr, citescore, citations, h-index</td>
<td>Scopus</td>
<td>Cron job</td>
</tr>
<tr>
<td>Fingerprints</td>
<td>Pure</td>
<td>Cron job</td>
</tr>
<tr>
<td>PubMed/ PubMedCentral ids</td>
<td>PubMed</td>
<td>Cron job</td>
</tr>
<tr>
<td>Impact Factor</td>
<td>Web of Science</td>
<td>Excel sheet import</td>
</tr>
<tr>
<td>citations, field weighted citation impact, author count</td>
<td>SciVal</td>
<td>Cron job</td>
</tr>
</tbody>
</table>
Advantages of aggregating data in Pure

• Display any content type on department web pages by exporting the data using Pure APIs

• Use Pure as a system of record for one or all content types

• Integrate Pure with other systems
  - Data visualization tools
  - Data warehouse
  - Faculty activity reporting systems
  - Patent databases (e.g.: Inteum)
  - Grant databases (e.g.: Cayuse)
# Bringing data into Pure

<table>
<thead>
<tr>
<th>Content Type</th>
<th>Auto-synchronization (.xml)</th>
<th>One-time legacy data bulk import (.xml)</th>
<th>Manual</th>
<th>Special uploads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons and Org units</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Masterlist</td>
</tr>
<tr>
<td>Publications and Patent data</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>- Profile Refinement Service for publications</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Automated Publication Imports</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Patent database connection (e.g: Inteum)</td>
</tr>
<tr>
<td>Grants &amp; Projects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>NIH Feed setup by Elsevier</td>
</tr>
<tr>
<td>Press/ Media</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Newsflo Feed setup by Elsevier</td>
</tr>
<tr>
<td>Equipment/ facilities</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Activities &amp; Prizes</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
## Assessing Automation

<table>
<thead>
<tr>
<th>Methodology</th>
<th>IT Requirements</th>
<th>Risks/ Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>None</td>
<td>Time and accuracy cost</td>
</tr>
<tr>
<td>Auto-synchronization (.xml)</td>
<td>IT resources</td>
<td>Setup time (depending on ease of ETL of data from IRs to Pure)</td>
</tr>
<tr>
<td></td>
<td>Xml expertise</td>
<td></td>
</tr>
<tr>
<td>One-time legacy data bulk import (.xml)</td>
<td>IT resources</td>
<td>Setup time (depending on ease of ETL of data from IRs to Pure)</td>
</tr>
<tr>
<td></td>
<td>Xml expertise</td>
<td></td>
</tr>
<tr>
<td>Special uploads</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Masterlist</td>
<td>None</td>
<td>Avg. maintenance time per year = 6 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual errors</td>
</tr>
<tr>
<td>• Scopus publication import feed (PRS)</td>
<td>None</td>
<td>PRS and Scopus data extraction fee depending on subscription</td>
</tr>
<tr>
<td>• Automated Import Candidates</td>
<td>None</td>
<td>None (depends on clients’ subscription to the sources)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual validation of content</td>
</tr>
<tr>
<td>• NIH Feed setup by Elsevier Projects + Grants Sync</td>
<td>None (enabled by Elsevier on request)</td>
<td>Coverage limited to NIH</td>
</tr>
<tr>
<td>• Newsflio</td>
<td>None (enabled by Elsevier on request)</td>
<td>Manual check required</td>
</tr>
</tbody>
</table>
Challenges with the Masterlist

Masterlist is used for adding basic HR information

- Limited granularity: Masterlist is a subset of the person/ org unit model
- Manual supplementation of profile related data

HR data is foundational content in Pure

- All related content requires a person and/or org unit in the data model
Challenges with the Masterlist

Scenario: Org restructuring/ New HR system on-campus
- New org hierarchy
- New org and person ids

Key aspect in Pure:
- Retaining content relations

Methodology:
- Script to map new ids to existing ids

Challenges:
- Manual re-work of entire Masterlist
- Subject to error
- Repercussions of manual errors cascade to other use-cases
Balancing automation and data quality in Pure

• Use Workflows in Pure to track any content updates in Pure
• Data clean-up jobs
  - Update Research Outputs regularly from online sources (v. 5.13)
• Bulk delete unused content
• Update org affiliation
Summary

• Benefits of automation:
  • No continuous manual intervention
  • Reduced human-errors
  • Frees up Admin time to focus on other system features and use cases

• Elsevier Support:
  • Process of moving from Masterlist to integration for clients that have already implemented Pure
  • More info on Expanding Pure
    • Day 2 - Thursday November 1st
    • 10:35-11:20
Thank you.
Appendix:
What are cron jobs in Pure?