Content Management for Content Enrichment: Architectural Issues and Strategies

Evan Owens
Chief Information Officer
AIP Publishing, LLC

STM E-Production Seminar 2013

This Presentation

- Historical Introduction
- Content Management & Content Enrichment
- Architectural Issues, Questions, Strategies, Use Cases

- AIP Publishing Case Study:
  - New Thesaurus
  - Author Disambiguation
  - Affiliation Disambiguation

A related presentation: *The Evolving Information Ecosystem of Publishing,* JATS-Con Proceedings 2010

http://www.ncbi.nlm.nih.gov/books/NBK48528

Publishing & Content Management in 1990s

"Publishing is adding a useful degree of uniformity to information"

How were we preparing for the digital future?

- Creating a version of record in SGML/XML full text
- Making the perfect master file
- Preparing to publish simultaneously to print and online
- Article SGML/XML file as a pseudo-database or pseudo-CMS:

```
<article copyeditor="XYZ" maildate="00/00/00">
```

- A document-centric approach!

Publishing & Content Management Today

"Publishing is adding value to a collection of content by enrichment and by managing the information life cycle"

What has changed?

- Content management is now multi-dimensional, multi-system
- Publishing is a much more complex ecosystem
  - DOIs, ORCID, DataCite, Thesauri, etc., etc., etc.
- Less static and less document-centric , more database-like
  - More complex information data models
- No longer "publish and be done"
  - Life cycle management is now essential component

Content Management

"A set of processes and technologies that support the evolutionary life cycle of digital information"

"Capture, storage, security, revision control, retrieval, distribution, preservation, and description of documents and content"

— Wikipedia 2007

Some CM components that support enrichment:

- Version control
- Technical metadata (formats, format versions, validation)
- Provenance metadata (processing history)

Questions for Enrichment Implementations

1. **Does the enrichment benefit from author vetting?**
2. **Is the enrichment part of the permanent scholarly record?**
3. **How standardized is the enriched information?**
4. **How volatile is the enriched information?**

Some Use Cases Examples

Use Cases

- Reference Linking
- Keywords
- Affiliations
- Authors Identity
- Funding Information

Key Questions:

Author vetting?

Scholarly record?

How standardized?

How volatile?

Key Architectural Choices

When is the content enhanced?

- By the author
- During submission
- During production/editorial process
- By the delivery/hosting system

Where does the enhanced information live?

- Embedded in the content
  - In the archival XML or in the exported XML
- External to the content
  - Layered information architectures

Key Design Challenges

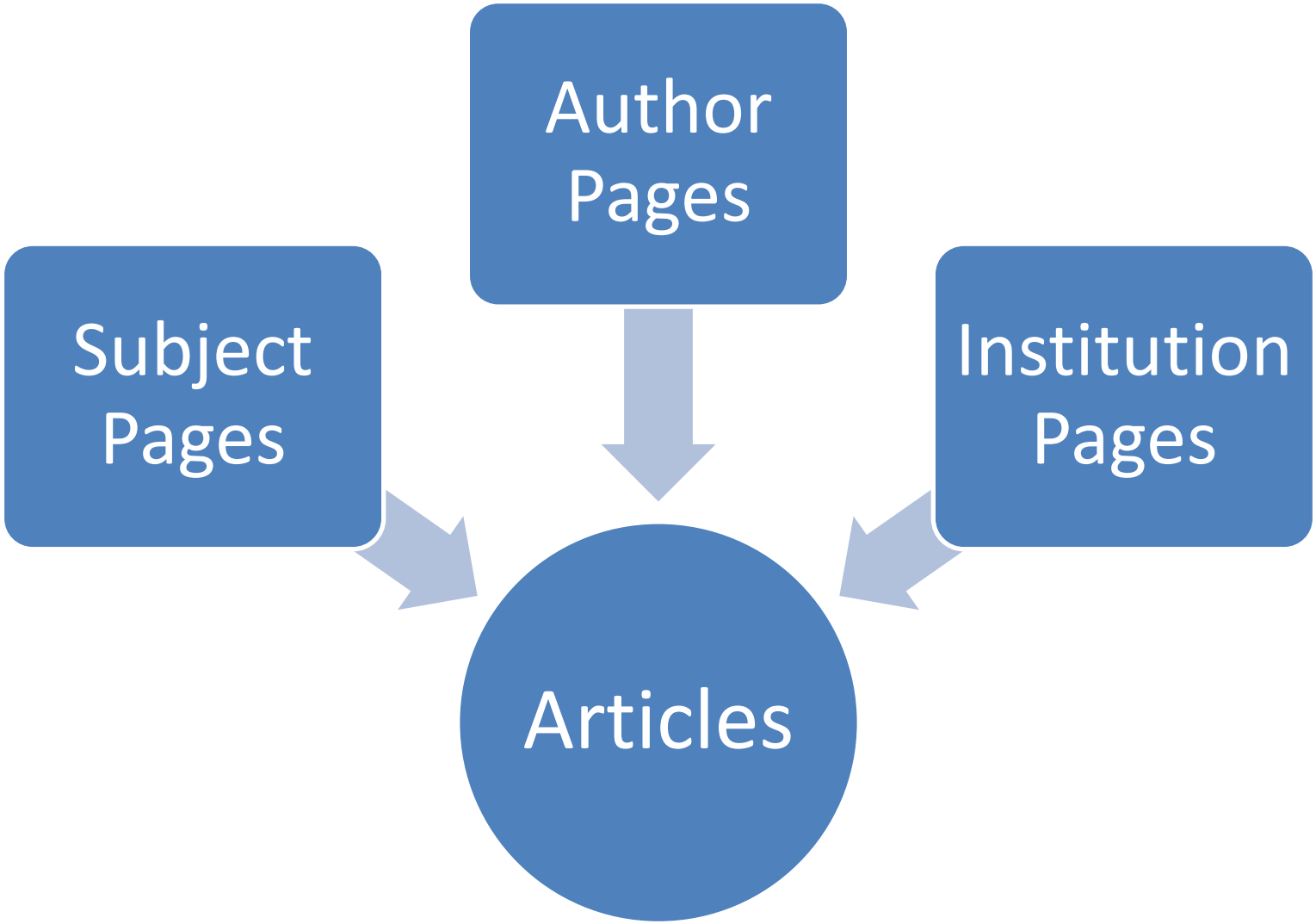
- What is the master source/copy of the information?
- Is the information normalized or de-normalized?
  - e.g., repeating parent metadata across child elements
- How to synchronize between multiple systems?
  - e.g., Peer Review System ⇔ XML content

As we move from document-centric to more complex information models and architectures, robust entity-relationship modeling becomes critical.

AIPP CASE STUDY:

NEW PHYSICS THESAURUS
AUTHOR DISAMBIGUATION
AFFILIATION DISAMBIGUATION

Enrichment / Disambiguation Goal



Key Strategic Decision (Business & Technical)

Semantic enrichment and disambiguation to be considered as a feature of the delivery platform

Not as part of the "publication" or "version of record"

Past practice:

- PACS codes printed on pages and in the PDF
- Resulted in mismatches between older and newer content
- Differences visible in previous hosting platform

Delivery/Hosting System Architecture

Publishing Technology's Pub2Web (P2W) hosting platform is built on an RDF triple store

- RDF is ideal for expressing complex relationships
- P2W manages RDF changes via set algebra
- P2W displays links dynamically based on the RDF

But

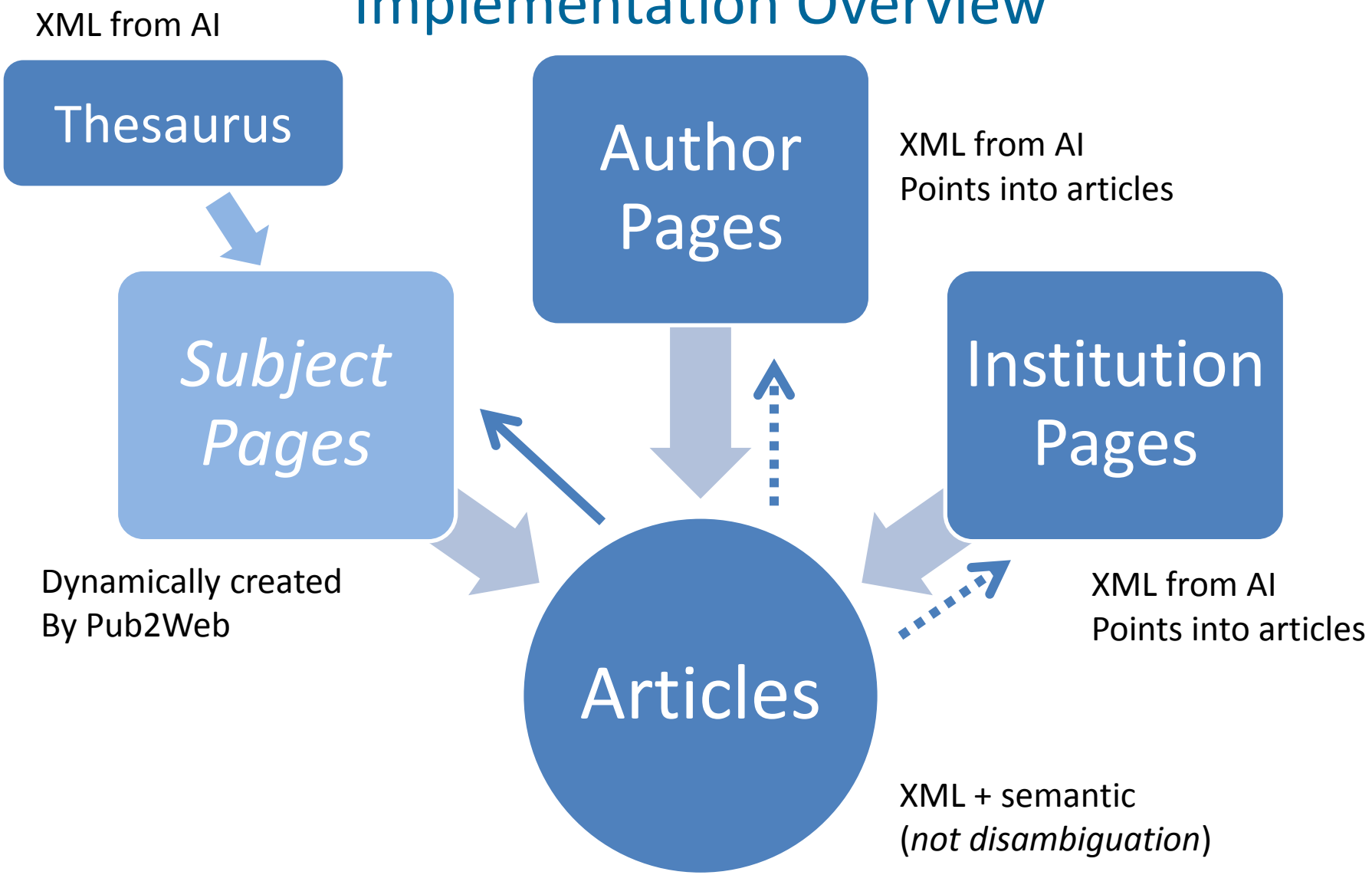
- Both parts of a relationship have to be present at RDF loading
- Content loading is very resource intensive

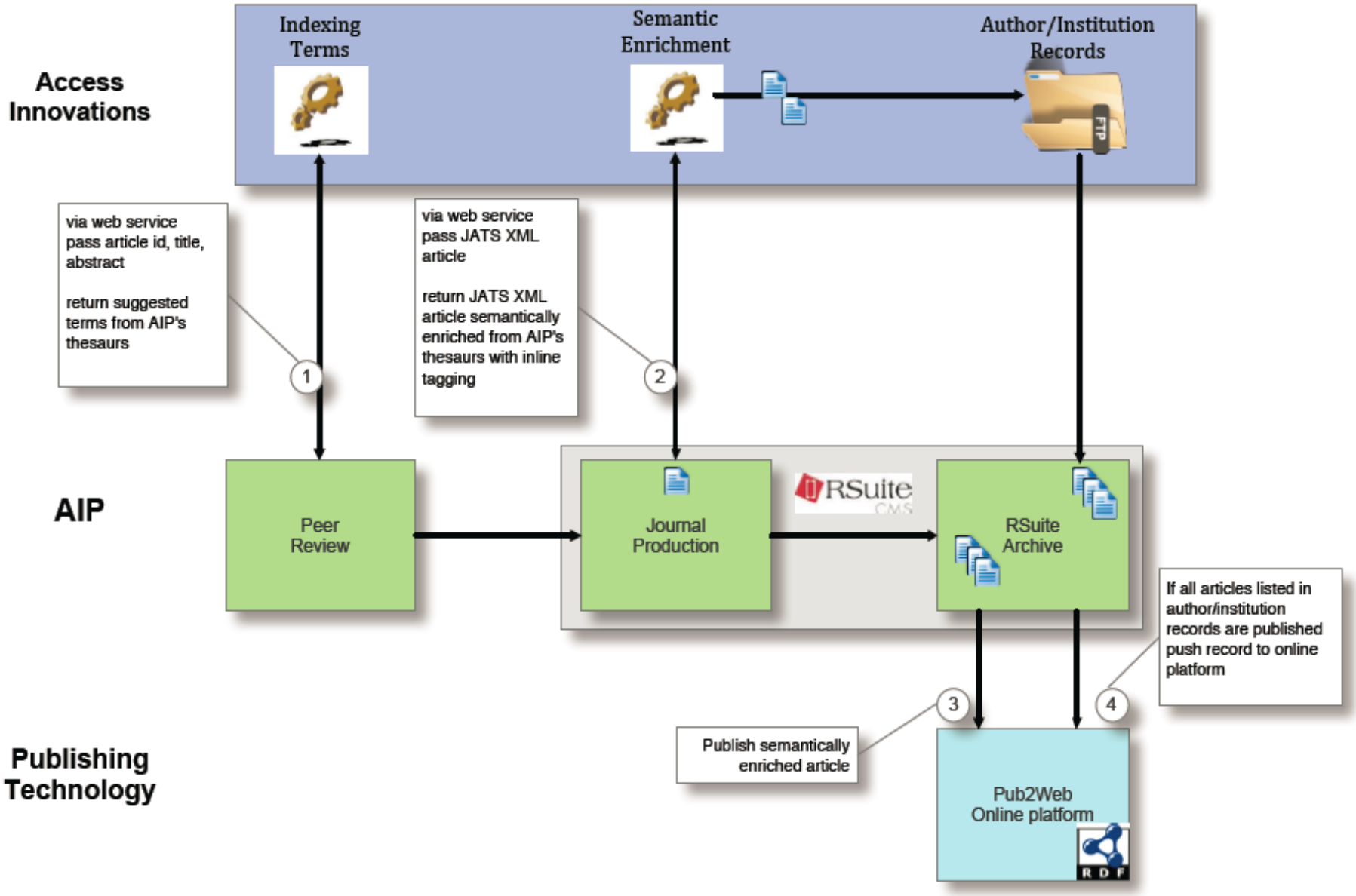
Every technology has its quirks ☺

AIPP's Implementation Choices

- **XML master in RSuite archive**
  - With all article assets: print, online, supplemental
  - Export packaging for hosting platform
  - Version control via RSuite
  - Interactions between AI and P2W managed by RSuite
  - Processing history captured in RSuite, not in the XML
- **Semantic embedded in article XML**
  - Keywords and inline tagging
  - XML markup strippable via XSLT
- **Disambiguation in separate XML files**
  - Pointing back into the articles
  - An external "annotation layer"

Implementation Overview




Access Innovations
Indexing Terms
Semantic Enrichment
Author/Institution Records
FTP
via web service pass article id, title, abstract
return suggested terms from AIP's thesaurs
1
via web service pass JATS XML article
return JATS XML article semantically enriched from AIP's thesaurs with inline tagging
2
AIP
Peer Review
Journal Production
RSuite CMS
RSuite Archive
If all articles listed in author/institution records are published push record to online platform
3
4
Publish semantically enriched article
Publishing Technology
Pub2Web Online platform
RDF

AIPP JATS XML: Keywords

- Keyword group in header:

```
<kwd-group kwd-group-type="sem:AIPThesaurus" specific-use="online">
<compound-kwd id="kwd1.1" content-type="sem:AIPTh1.2">
<compound-kwd-part content-type="value">Free energy</compound-kwd-part>
<compound-kwd-part content-type="code">2237</compound-kwd-part>
<compound-kwd-part content-type="weight">34</compound-kwd-part>
</compound-kwd>
<compound-kwd id="kwd1.2" content-type="sem:AIPTh1.2">
<compound-kwd-part content-type="value">Boundary value problems</compound-kwd-part>
<compound-kwd-part content-type="code">709</compound-kwd-part>
<compound-kwd-part content-type="weight">27</compound-kwd-part>
</compound-kwd>
</kwd-group>
```

- Keywords inline:

```
In terms of the thermodynamic limit of the specific
<named-content content-type="sem:AIPTh1.2" rid="kwd1.1">free energy</named-content>
the same quantity can be expressed as ...
```

AIPP Disambiguated Author XML

```
<author-id source="SciAuth">AU0773954</author-id>
<name surnamefirst="no"><surname>Aizenman</surname><given-names>Michael</given-names></name>
<affiliations><aff-id>AF0000377</aff-id></affiliations>
<publications>
    <publication contrib-type="author">
        <article-id pub-id-type="doi">10.1063/1.3679069</article-id>
        <contrib-seq>1</contrib-seq>
    </publication>
    ...
</publications>
<kwd-group kwd-group-type="sem:AIPThesaurus">
    <compound-kwd content-type="sem:AIPTh1.2">
        <compound-kwd-value>Free energy</compound-kwd-value>
        <compound-kwd-code>2237</compound-kwd-code>
        <compound-kwd-weight>34</compound-kwd-weight>
    </compound-kwd>
    ...
</kwd-group>
<coauthors>...</coauthors>
<emails>...</emails>
```

AIPP Disambiguated Affiliation XML

```
<aff-id>AF0000377</aff-id>
<institution-name>Princeton University</institution-name>
<departments>
    <department>
        <department-name>Department of chemistry</department-name>
        <address/>
    </department>
    <department>
        <department-name>Princeton Institute for the Science and Technology
         of Materials (PRISM)</department-name>
        <address/>
    </department>
    ...
</departments>
<publications>
    <publication>
        <article-id pub-id-type="doi">10.1063/1.555605</article-id><aff-seq>2</aff-seq>
    </publication>
    <publication>
        <article-id pub-id-type="doi">10.1063/1.3555832</article-id><aff-seq>1</aff-seq>
    </publication>
    ...
</publications>
```

AIPP Lifecycle Use Cases: Add / Change / Delete

- Content
  - Corrections could impact author / affiliation / keywords
- Thesaurus Terms
  - Vocabulary will evolve
- Enrichment Rules
  - Quarterly reruns of entire corpus with latest rule set
- Author Disambiguation
  - New info could cause merge or split
- Institution Disambiguation
  - Organizational changes (names, mergers, etc.)

"Publishing is adding value to a collection of content by enrichment and by managing the information life cycle"

NO SINGLE MAGIC SOLUTION
YOUR MILEAGE MAY VARY!

QUESTIONS? COMMENTS?