8-1-2013

Let There Be Light! Indexing Items from Digital Commons in Apache Solr via OAI-PMH

Graham Hukill
Wayne State University, ej2929@wayne.edu

Let There Be Light!

Indexing Materials from Digital Commons in Apache Solr with OAI-PMH

Graham Hukill
Digital Publishing Librarian
graham.hukill@wayne.edu

Solr is a powerful, fast, open source, enterprise level, **full-text search platform**, based on the Lucene Index. Solr is becoming a common piece of infrastructure in libraries and digital collections infrastructure.

Documents are added to Solr by "indexing" Solr ready XML files which contain the text will be made searchable.

Finally, **Searching** is done through a separate interfaces that pulls in Solr search results.

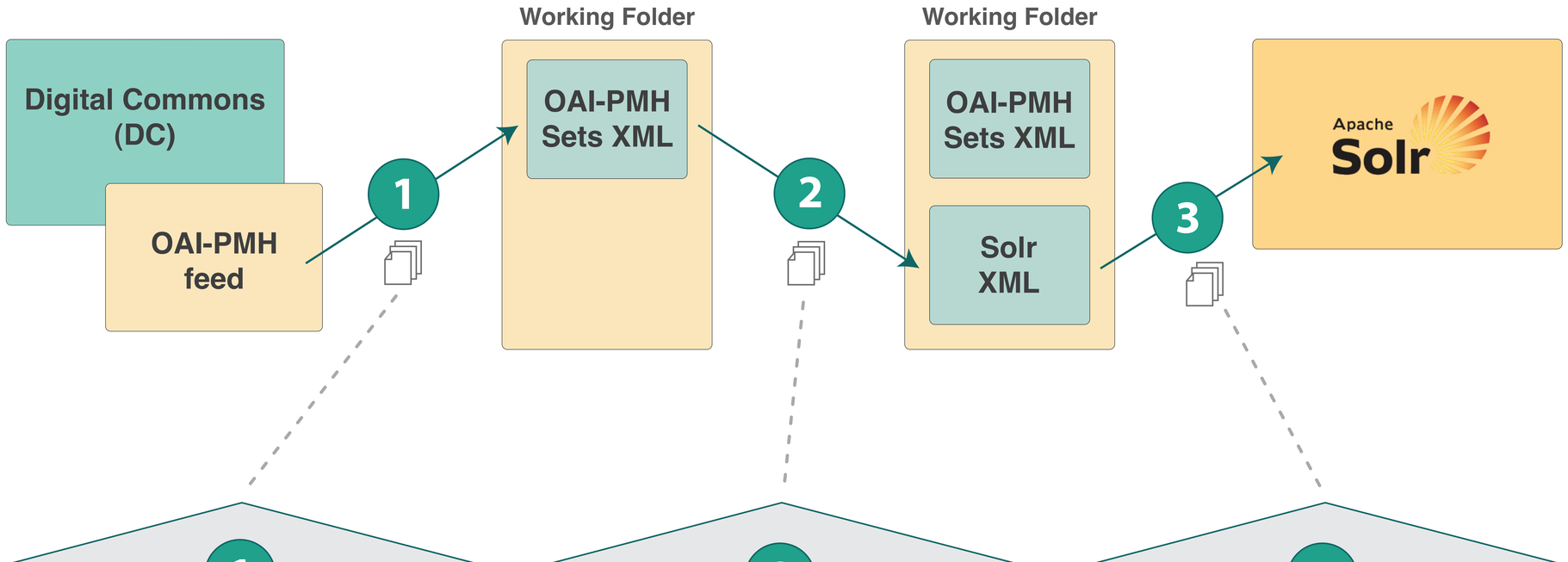
lucene.apache.org/solr

OAI-PMH

Open Archive Initiative - Protocol for Metadata Harvesting (**OAI-PMH**) is a widely used protocol for sharing structured metadata, in **XML** form, between repositories. Digital Commons (DC) provides metadata records for all items in an institution's repository via an "OAI-PMH feed."

Each **Series** in DC comes through as a **Set** in the OAI-PMH feed. Each **Item** in DC is expressed as a **Record** within a Set. It is these **Records** that are indexed in Solr.

www.openarchives.org/pmh



1

Download OAI-PMH Set XML from Digital Commons (DC)

- DC's OAI-PMH feed provides 100 Records at a time, download in 100 record chunks
- Each Record contains **Dublin Core** metadata for each item (title, description, subject, date, etc.)
- Slowest part of the process...

2

Convert OAI-PMH XML to Solr ready XML

- Information is indexed in Solr with documents that contain all text that will be searchable, in a format easily indexed by Solr. Below is an example...

```
<add>
   <doc>
      <field type="title">Title Here</field>
      <field type="subject">biology</field>
   </doc>
<doc>
      <field type="title">Another Title Here</field>
      <field type="subject">music</field>
   </doc>
</add>
```

3

Index Records in Solr

- Now that the records have been converted from OAI-PMH XML to Solr XML, we can index them in Solr
- We are doing this nightly at midnight
- Currently, we overwrite all Solr records at each nightly indexing, also adding new ones at that time. Next iterations could leverage DC OAI datestamp to only index modified records.

This utility is available on GitHub:
github.com/WSULib/dc2Solr

Why bother?

- Fully automated, takes approximately **13 minutes for 3,000+ records**
- Potentially pulls Digital Commons records into Library **discovery system** via Solr
- Provides ability to **audit** records from Digital Commons by aggregating them in one place.