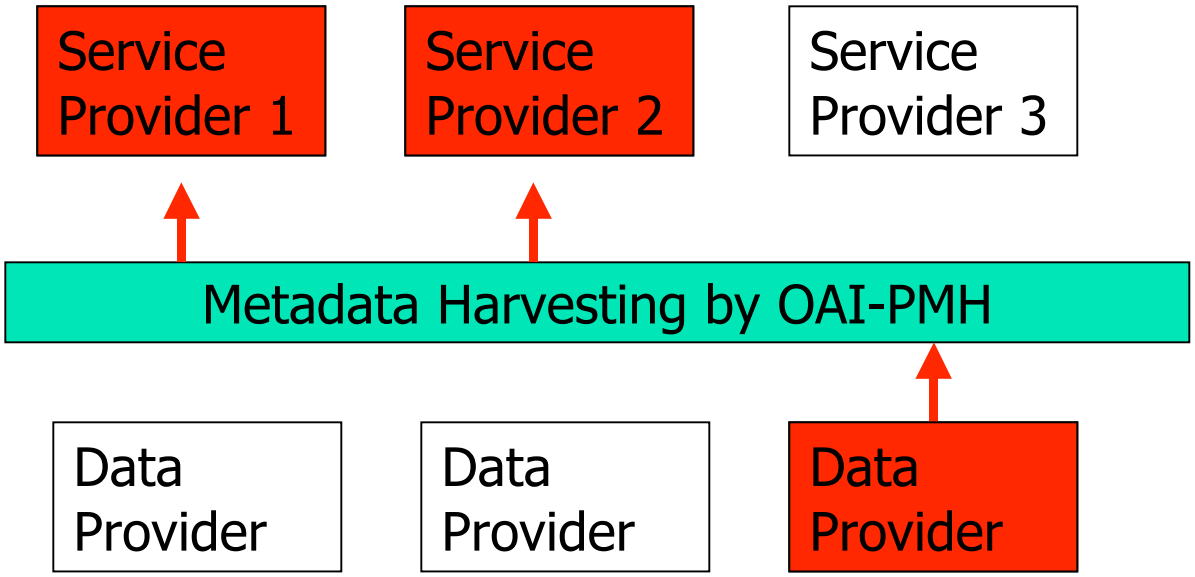
Repository Synchronization in the OAI Framework

Xiaoming Liu
DL Research and Prototyping
Los Alamos National Laboratory

OAI Framework and Synchronization Problem



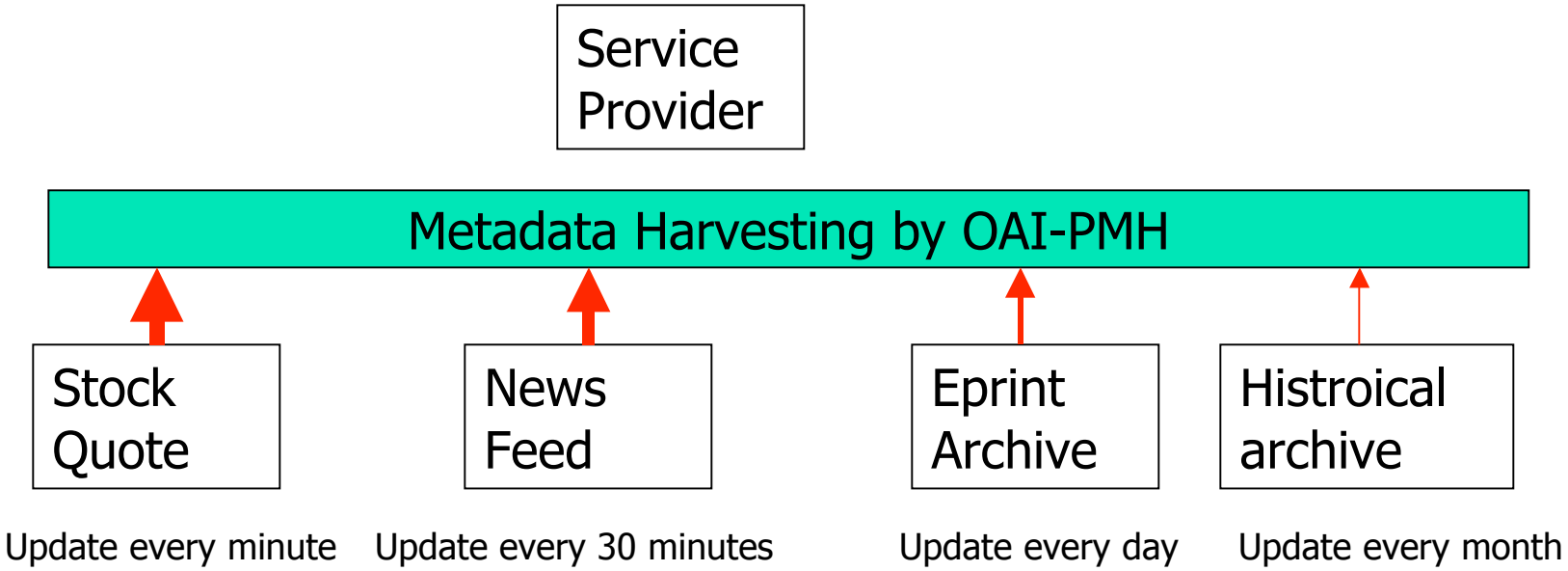
- Service Provider periodically polls data providers for new data.

Why important?

- Michael Nelson in 2nd Workshop on the Open Archives Initiative.

  "Premise: ***OAI-PMH is applicable to any scenario that needs to update / synchronize distributed state.*** Future opportunities are possible by creatively interpreting the OAI-PMH data model"
- Possible scenarios
  - Large number of data providers and service providers.
  - Annotation, review services, log files expose in DL applications.
  - Other applications, such as stock quote and news aggregation.

Example


Service Provider
Metadata Harvesting by OAI-PMH
Stock Quote
News Feed
Eprint Archive
Histroical archive
Update every minute
Update every 30 minutes
Update every day
Update every month

Experiments

- Arc (http://arc.cs.odu.edu) harvester. Till May, 2003, Arc collected ~6.5M records from 162 data providers.
- the result of this paper is based on period 09/2001 – 09/2002 with about 100 data providers.
- the change rate includes new, modified, and deleted records.
- we observe the update rate and update interval.

Update Frequency of Data Providers

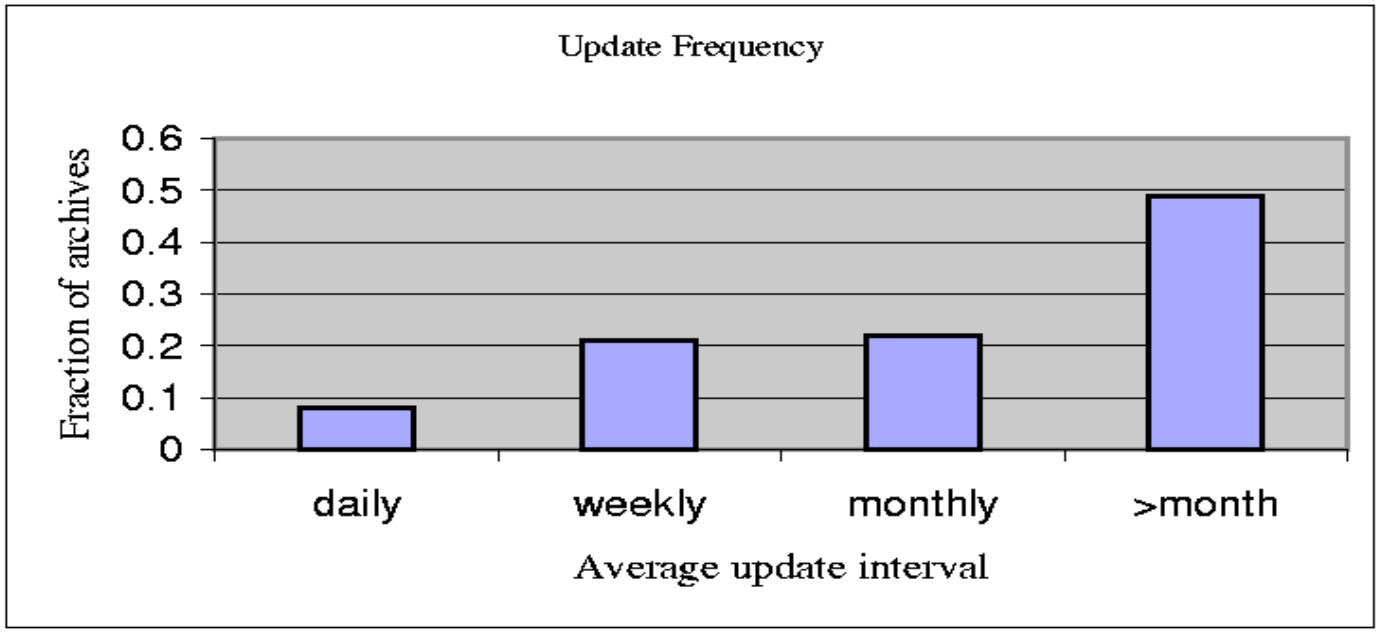


Figure 4.2: Average update frequencies of OAI-PMH repositories

- The update interval varies dramatically from site to site.

Trend of Update Frequency

- Many data providers change at a constant rate overall.
  - E-print type repositories have a small but steady stream of ongoing daily or weekly updates.
  - Museum or historically oriented archives have an initial burst period of accession (perhaps all at once), but then trickle down to just infrequent changes.
- The update frequency varies dramatically from site to site.

Approaches to Improve Freshness

Approaches to Improve Freshness

- Inside OAI-PMH.
  - Best estimation.
    - Harvester estimates the update frequency by learning the harvest history.
  - Syndication.
    - Data provider describes its update frequency explicitly.
- Beyond OAI-PMH.
  - Subscribe/notify.
    - Data providers notify a service provider whenever their content is changed.
  - Push model.
    - Data providers directly push updates to service provider.

Best Estimation

- The harvester estimates the record update frequency by learning the harvest history.
- A harvester may not necessarily provide 100% freshness at any time, for example, a harvester may harvest repositories with higher average update frequency more frequently, and harvest all other repositories once a week.

Syndication Container

- A data provider may describe its update frequency in an optional container of OAI-PMH Identify response.
- RSS (Rich Site Summary)
  - UpdatePeriod (Describes the period over which the data provider is updated),
  - UpdateFrequency (Describe the frequency of updates in relation to the update period)
  - UpdateBase (Defines a base date to be used in concert with updatePeriod and updateFrequency.

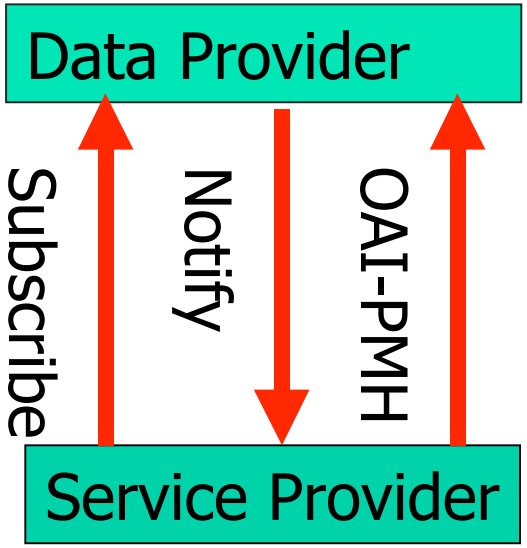
XML Schema for Syndication

```
<?xml version="1.0" encoding="UTF-8"?>
<schema targetNamespace="http://purl.org/rss/1.0/modules/syndication/"
  xmlns="http://www.w3.org/2001/XMLSchema"
  xmlns:syndication="http://purl.org/rss/1.0/modules/syndication/"
  elementFormDefault="qualified" attributeFormDefault="unqualified">
 <element name="syndication">
    <complexType>
      <sequence>
        <element name="updatePeriod" minOccurs="0" maxOccurs="1"
                 type="syndication:updatePeriodType"/>
        <element name="updateFrequency" minOccurs="0" maxOccurs="1"
                 type="integer"/>
        <element name="updateBase" minOccurs="0" maxOccurs="1"
                 type="dateTime"/>
      </sequence>
    </complexType>
  </element>
<simpleType name="updatePeriodType">
    <restriction base="string">
      <enumeration value="hourly"/>
      <enumeration value="daily"/>
      <enumeration value="weekly"/>
      <enumeration value="monthly"/>
      <enumeration value="yearly"/>
    </restriction>
  </simpleType>
</schema>
```

XML Sample for syndication container

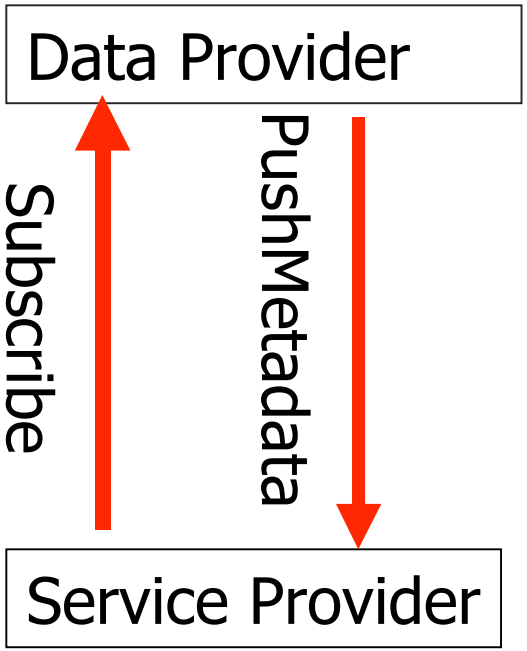
```
<description>
- <syndication xmlns="http://purl.org/rss/1.0/modules/syndicatoin/"
    xmlns:xsi="http://purl.org/rss/1.0/modules/syndicatoin/"
    xsi:schemaLocation="http://dlib.cs.odu.edu/OAI/2.0/syndication
    http://dlib.cs.odu.edu/OAI/2.0/syndication.xsd">
    <updatePeriod>hourly</updatePeriod>
    <updateFrequency>2</updateFrequency>
    <updateBase>1999-02-01T00:00</updateBase>
  </syndication>
</description>
```

Subscribe/Notify model



- Advantage: Useful for a data provider with irregular update frequency.
- Disadvantage:
  - A service provider needs to listen for "notify" signal.
  - A data provider needs keep a list of subscribed service providers.
  - Beyond OAI.

Push Model



- Advantage:
  - Useful for a data provider with irregular update frequency.
  - Bypass NAT/firewall
- Disadvantage:
  - A service provider needs to listen for “pushmetadata” requests.
  - Beyond OAI.

Proposed Work to OAI Community

- Investigate the freshness problem.
- Add syndication container as an optional container in "*Identify*" response (Implementation guideline). This can be based on the RSS syndication format.
- Investigate the community for the requirement of "subscribe" and "push" model.