ITHAKA

JSTOR | PORTICO | ITHAKA S+R

Beyond Well-Formed and Valid

QA for XML Configuration Files

By Sheila M. Morrissey
Senior Research Developer, Portico

Balisage 2012
International Symposium on Quality Assurance and Quality Control in XML
August 06, 2012

Rich XML Configuration Files

ITHAKA

- **Ithaka S+R** is a research and consulting service that focuses on the transformation of scholarship and teaching in an online environment, with the goal of identifying the critical issues facing our community and acting as a catalyst for change.

- **JSTOR** is a research platform that enables discovery, access, and preservation of scholarly content.

- **Portico** is a digital preservation service for e-journals, e-books, and other scholarly e-content.

Portico - Third Party Preservation

PORTICO

Portico is among the largest community-supported digital archives in the world.

Working with libraries, publishers, and funders, we preserve e-journals, e-books, and other electronic scholarly content to ensure researchers and students will have access to it in the future.

PORTICO

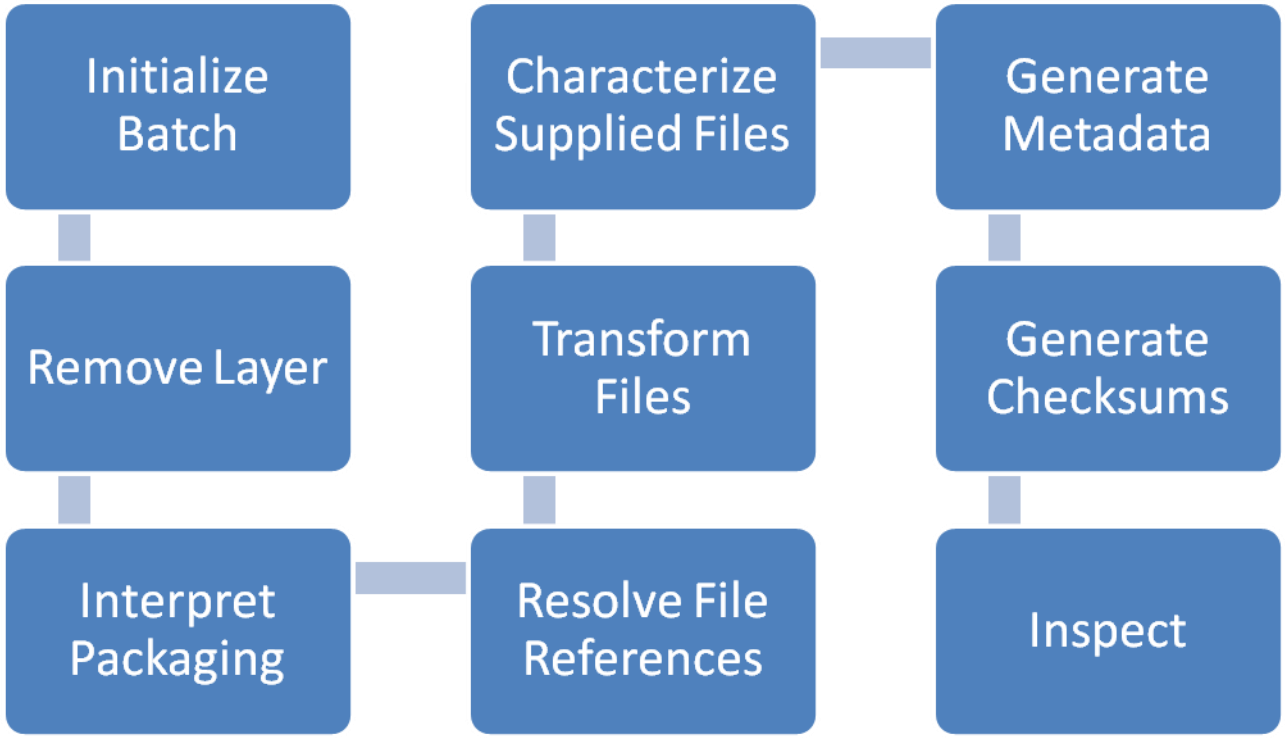
Portico – Preserved Content

PORTICO

Preserved Content

» E-journal titles	9,568
» E-book titles	16,861
» D-collections	12
» Archival Units	19,433,869
» Preserved Files	319,737,011

Portico XML Configuration: Format-Driven Workflow



PORTICO

Portico XML Configuration: Packages & Profiles

Incoming File System

- PublisherA
  - 0008543x
    - 2006
      - 106
        - B
          - CNCR21779
            - 21779_ftp.pdf
            - 21779_ftp.sgm
            - equation
              - aueq001.tif
              - aueq002.tif
              - nueq001.gif
              - nueq002.gif
            - image_m
              - mfig001.jpg
            - image_n
              - nfig001.jpg
            - image_t
              - tfig001.gif

Portico XML Configuration: Packages & Profiles

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

Incoming File System

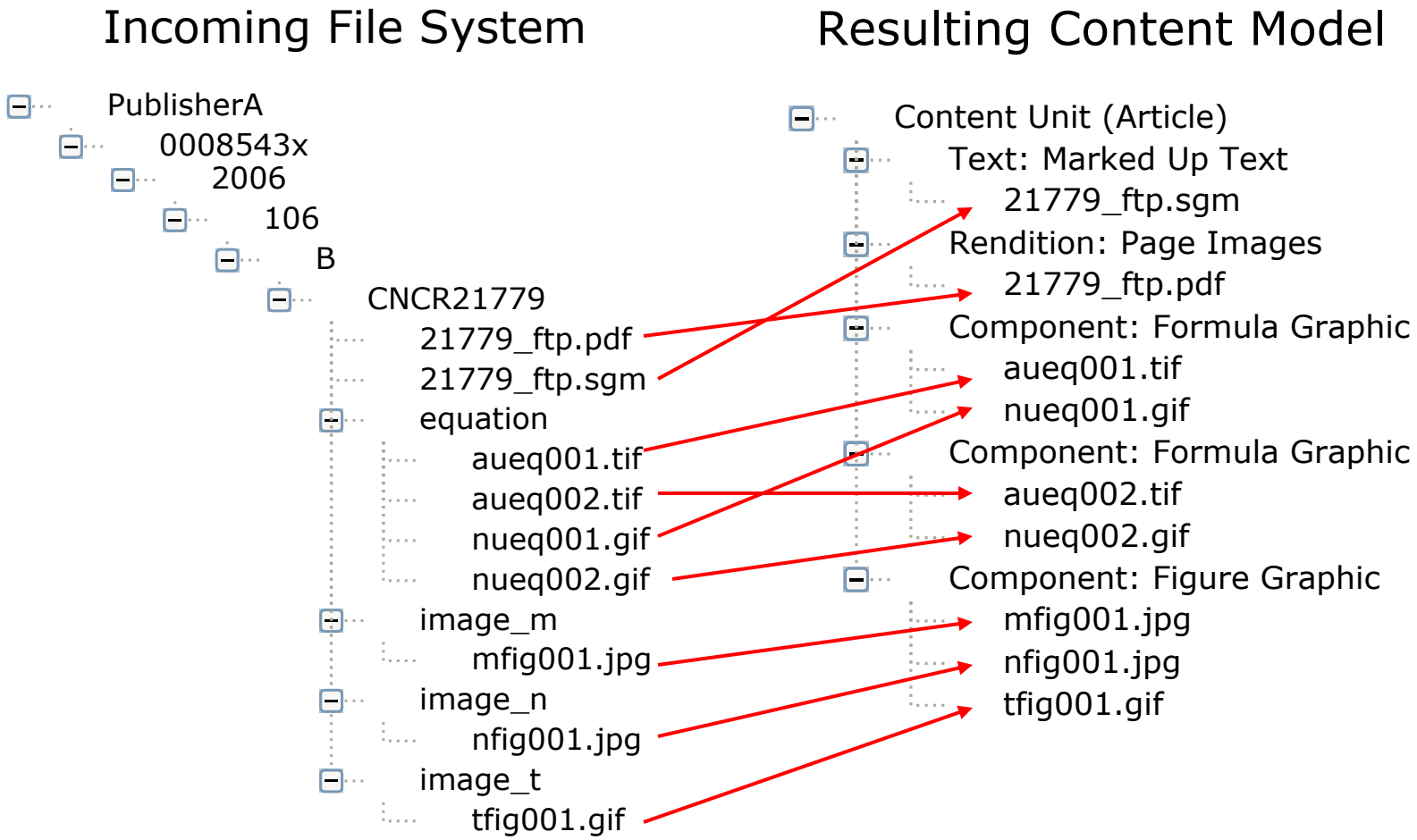
- PublisherA
  - 0008543x
    - 2006
      - 106
        - B
          - CNCR21779
            - 21779_ftp.pdf
            - 21779_ftp.sgm
            - equation
              - aueq001.tif
              - aueq002.tif
              - nueq001.gif
              - nueq002.gif
            - image_m
              - mfig001.jpg
            - image_n
              - nfig001.jpg
            - image_t
              - tfig001.gif

Resulting Content Model

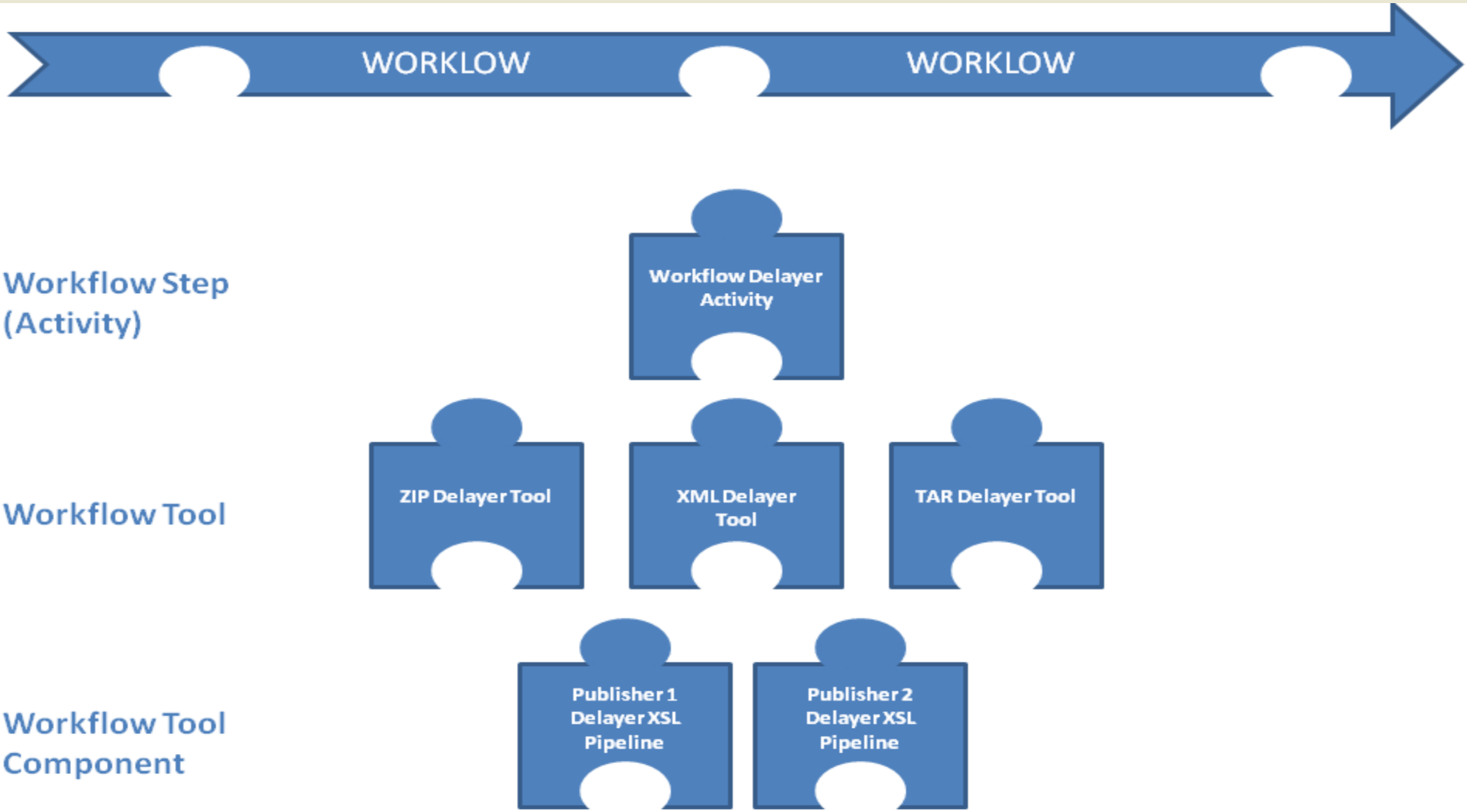
- Content Unit (Article)
  - Text: Marked Up Text
    - 21779_ftp.sgm
  - Rendition: Page Images
    - 21779_ftp.pdf
  - Component: Formula Graphic
    - aueq001.tif
    - nueq001.gif
  - Component: Formula Graphic
    - aueq002.tif
    - nueq002.gif
  - Component: Figure Graphic
    - mfig001.jpg
    - nfig001.jpg
    - tfig001.gif

PORTICO

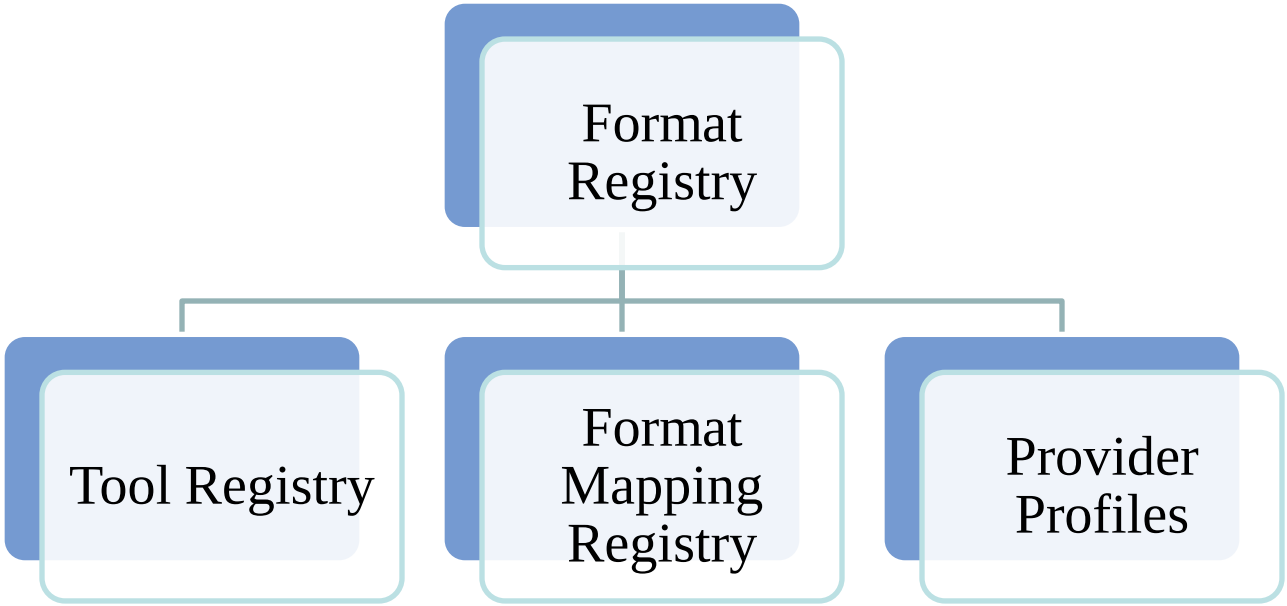
Portico XML Configuration: Packages & Profiles



Portico XML Configuration: Format-Driven Workflow



Format-Driven Workflow: Format Registry



Format-Driven Workflow: Format Registry


```
<Format FormatId=
      "XXX_NLM_Journal_Publishing_DTD_2.1"
         CreationTimestamp=
      "2006-06-13T13:00:00-05:00">
  <PorticoDefinedName>
    XXX Journal Publishing DTD v2.1 20050630
  </PorticoDefinedName>
    …
<Format>
```

Format-Driven Workflow: Tool Registry

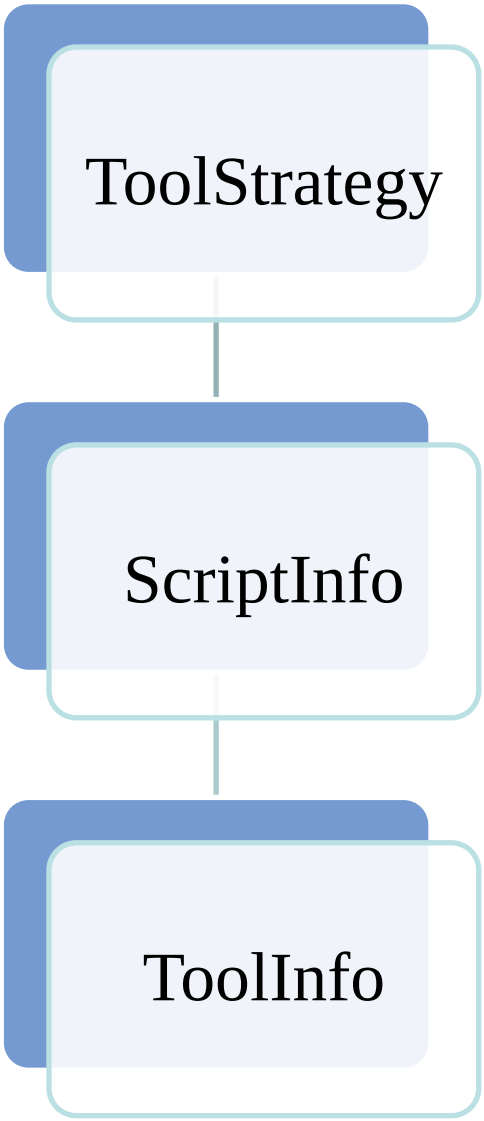
ITHAKA
JSTOR | PORTICO | ITHAKA S+R

```
<TransformationSet>
    <ToolStrategy
      SupportingFormatId=
      "XXX_NLM_Journal_Publishing_DTD_2.1">
       <Script Rid="scrxxx"/>

    </ToolStrategy>
    ...
</TransformationSet>
```


PORTICO

Format-Driven Workflow: Tool Registry



PORTICO

Format-Driven Workflow: Tool Registry


```
<ScriptInfo ScriptId="scrxxx”>
   <Tool Rid="BaseTransform_1.0">
    <Parameters>
     <Parameter>
      <Name>StyleSheetList</Name>
      <ValueOrderedList>
        <ValueOrderedListItem>
          <Number>10</Number>
          <Value>xxx2ptc_1_10.xsl</Value>
        </ValueOrderedListItem>
      </ValueOrderedList>
    </Parameter>
    ....
   </Parameters>
  </Tool>
 </ScriptInfo>
```

PORTICO

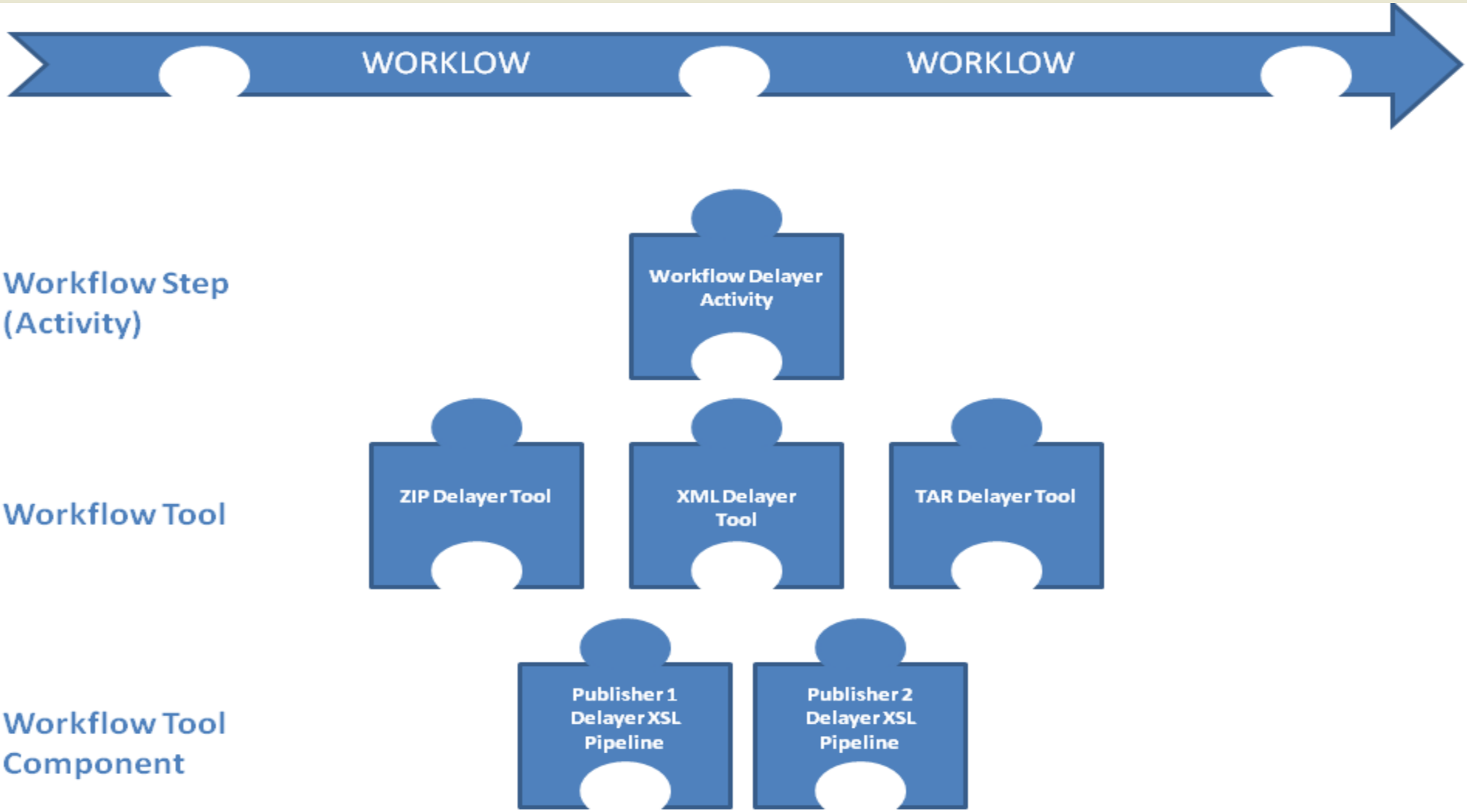
Format-Driven Workflow: Tool Registry

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

```
<ToolInfo Id="BaseTransform_1.0">
 <Name>BaseTransformTool:1.0:2007-05-
  01</Name>
 <Description>Tool for transformation of XML
  files via XSL stylesheets. </Description>
 <Status>ACTIVE</Status>
 <ClassName>
org.portico.threadedtool.tool.transform.Base
  TransformTool
  </ClassName>
 <ToolParameters>
   <ToolParameter Name="StyleSheetList"
      Required="true"
     ParameterType="ValueOrderedList">
<Description>
</ToolParameters> …
```

PORTICO

Portico XML Configuration: Format-Driven Workflow



Portico XML Configuration: Deployment Issues

Consistency

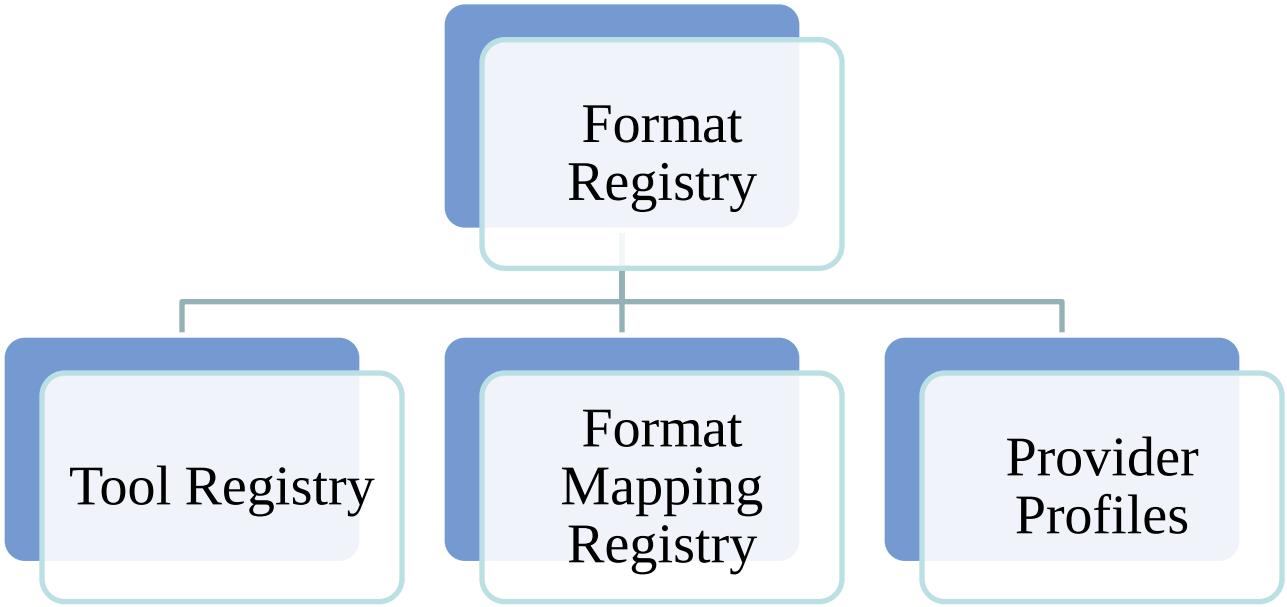
```
<Parameter>
 <Name>InputFilterClass</Name>
 <Value>
 org.portico.threadedtool.tool.transform.filter.ExternalEntityReplacerFilter
 </Value>
 <Parameter>
  <Name>AttributeValueSeparator</Name>
  <Value>:</Value>
</Parameter>
</Parameter
```

```
public void initialize(ToolRequest request) throws Exception {
parmValue =
request.getObjectArgument(IToolConstants.ARG_TOOL_ATTRIBUTEVALUESEPARATOR);
if (parmValue != null){
        if (parmValue instanceof String) {
                this.setAttributeValueSeparator((String)parmValue);}
                else {
                throw new Exception("");}
}
```

Portico XML Configuration: Deployment Issues

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

- "Referential Integrity"

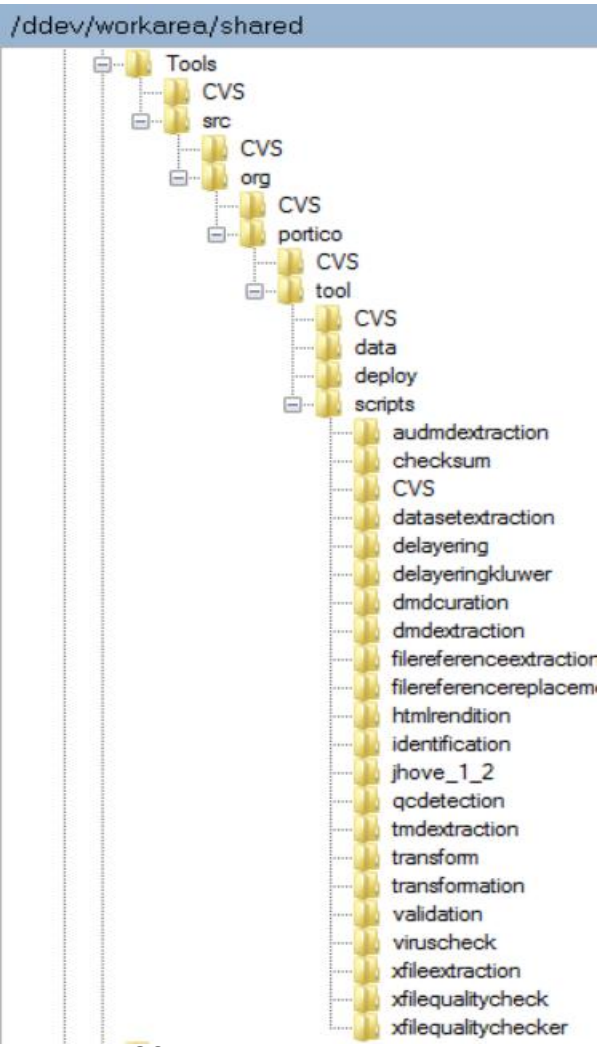


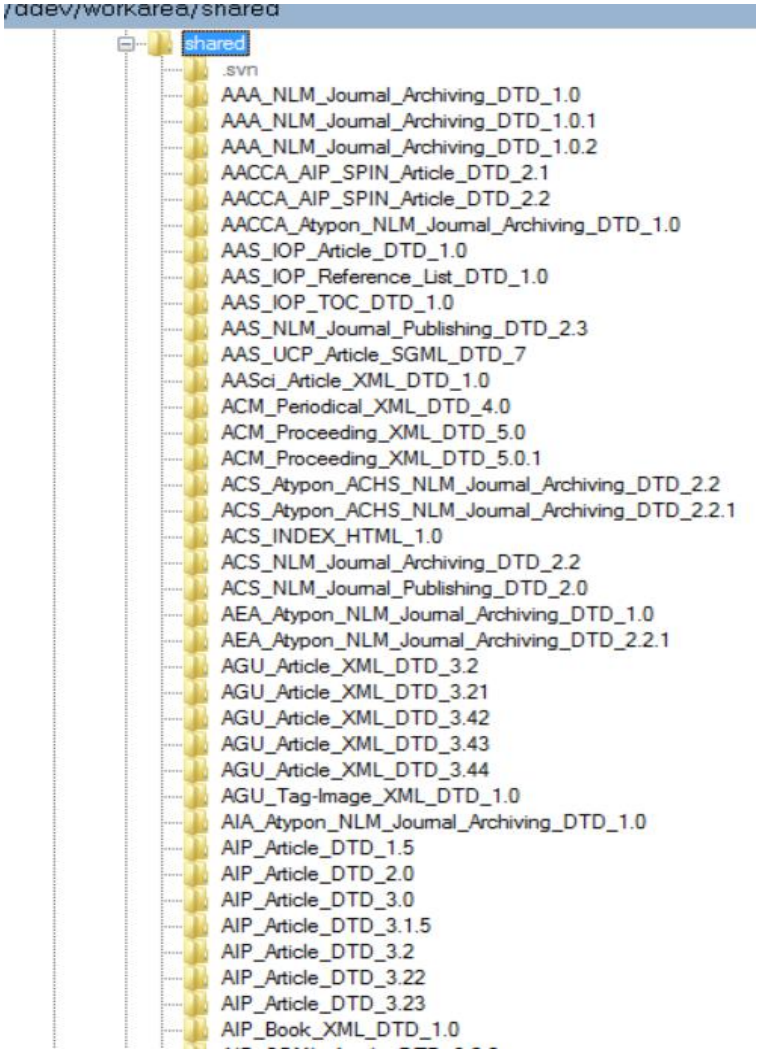
PORTICO

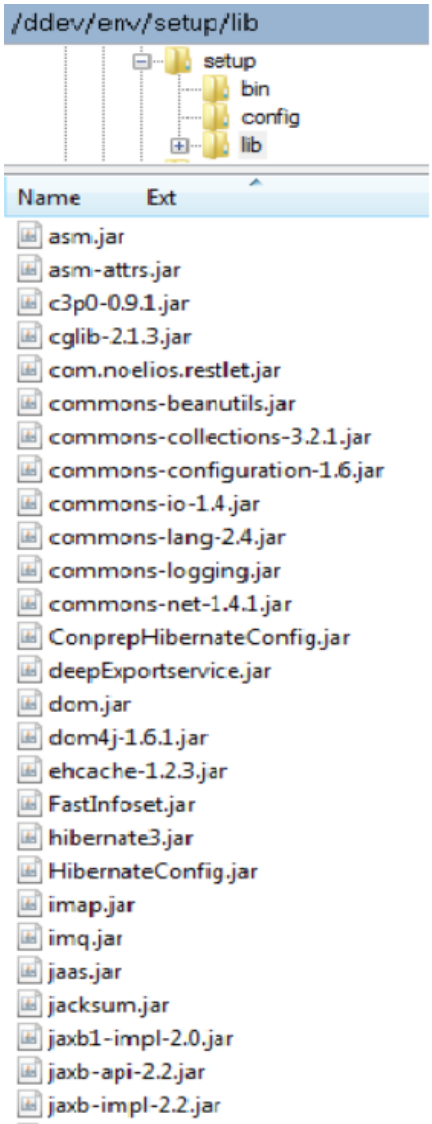
Portico XML Configuration: Deployment Issues

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

❖ Existence







Deployment Solutions: Consistency QA


```
<ToolInfo
  Id="BaseTransform_1.0">
 …
 <ToolParameters>
   <ToolParameter
  Name="AttributeValueSeparator"
       Required="false"
      ParameterType="Value">
<Description>
</ToolParameters> …
```

Deployment Solutions : “Referential Integrity” QA


```
<xsl:template match="@SupportingFormatId" mode="formatId-
   check">
<xsl:param name="regDirUri"/>
<xsl:variable name="formatId" select="."/>
<xsl:variable name="formatIdExists"
   select="frf:formatIdExists($jRegExtFunc, $formatId,
   $regDirUri)"/>
<xsl:if test="not($formatIdExists)">
    <xsl:variable name="pathToNode">
   <xsl:apply-templates select="." mode="get-full-
   xpath"/>
    </xsl:variable>
    <xsl:message terminate="no">Format id <xsl:value-of
   select="."/> at node <xsl:value-of
   select="$pathToNode"/> does not exist in the Format
   Registry.</xsl:message>
</xsl:if>
</xsl:template>
```

Deployment Solutions : Existence QA


```
<!-- ========================================================= -->
<!--              name="check-file-exists"                     -->
<!-- ========================================================= -->
<xsl:template name="check-file-exists">
<xsl:param name="scriptType"/>
<xsl:param name="scriptDir"/>
<xsl:param name="fileName"/>
<xsl:param name="Path"/>
<xsl:variable name="result"
        select="jexist:fileExistsOnClasspath($jExistInstance,
                  $scriptType, $scriptDir, $fileName)"/>
<xsl:if test="not($result)">
    <xsl:message  terminate="no" >
       File  <xsl:value-of select="$fileName"/> does not exist.
       ScriptType = <xsl:value-of select="$scriptType"/>,
       ScriptDir = <xsl:value-of select="$scriptDir"/>
       Path to script = <xsl:value-of select="$Path"/>
    </xsl:message>
</xsl:if>
</xsl:template>
```

XML Config QA: Reflections

Choices

- Properties files
- Relational DB

PORTICO

XML Config QA: Reflections

Limits

- Manual co-ordination
- Deploy time vs. Run time existence

XML Config QA: Reflections

Aesthetics

Image: Wikimedia Commons (http://commons.wikimedia.org/wiki/File:River_terrapin.jpg)

PORTICO

Questions or Comments?

Sheila Morrissey
sheila.morrissey@ithaka.org
@sheilaMorr
www.portico.org