

I T H A K A

JSTOR | PORTICO | ITHAKA S+R

# Beyond Well-Formed and Valid

QA for XML Configuration Files

By Sheila M. Morrissey

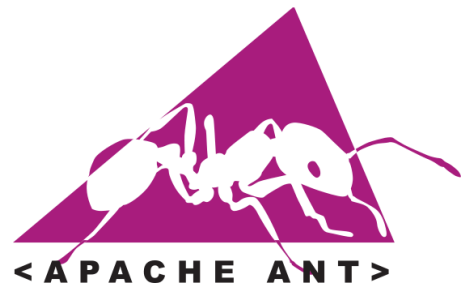
Senior Research Developer, Portico

Balisage 2012

International Symposium on Quality  
Assurance and Quality Control in XML

August 06, 2012

# Rich XML Configuration Files





- **Ithaka 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

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



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.



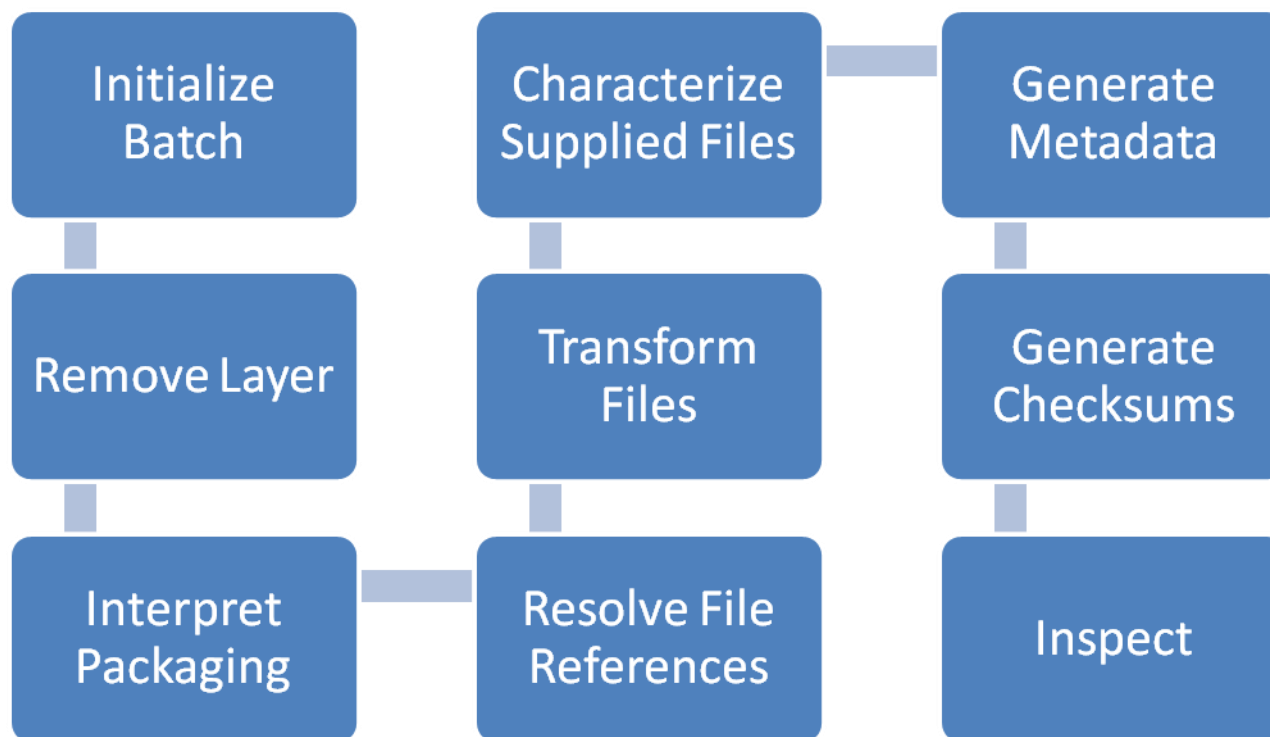
PORTICO

## 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

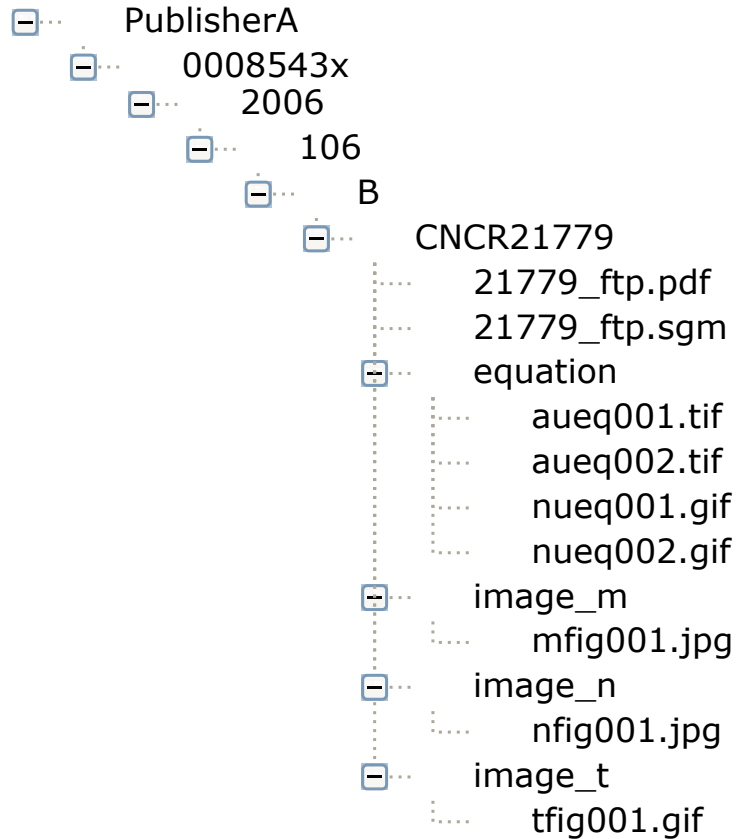


# Portico XML Configuration: Format-Driven Workflow



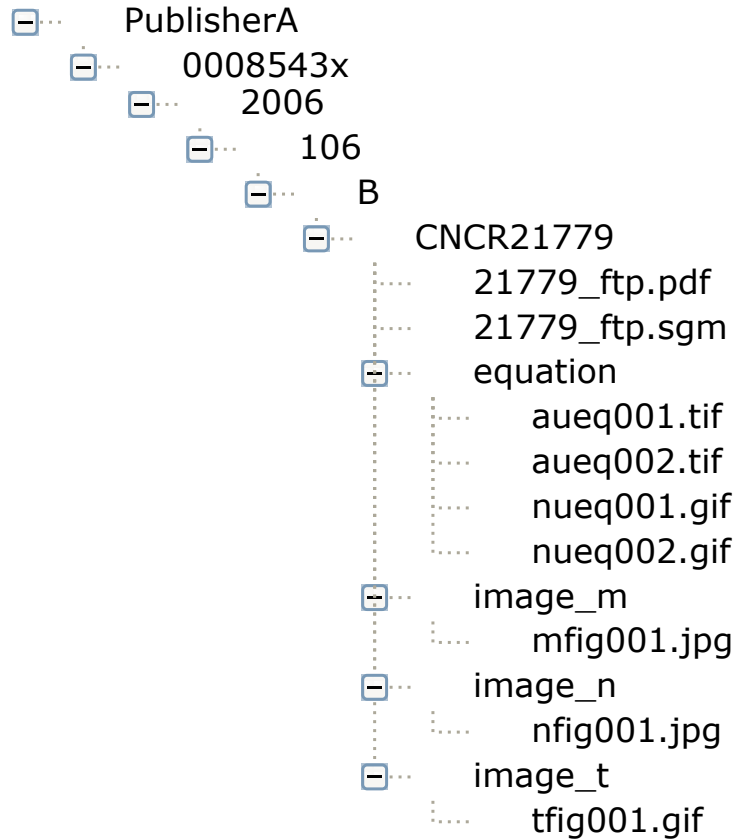
# Portico XML Configuration: Packages & Profiles

## Incoming File System

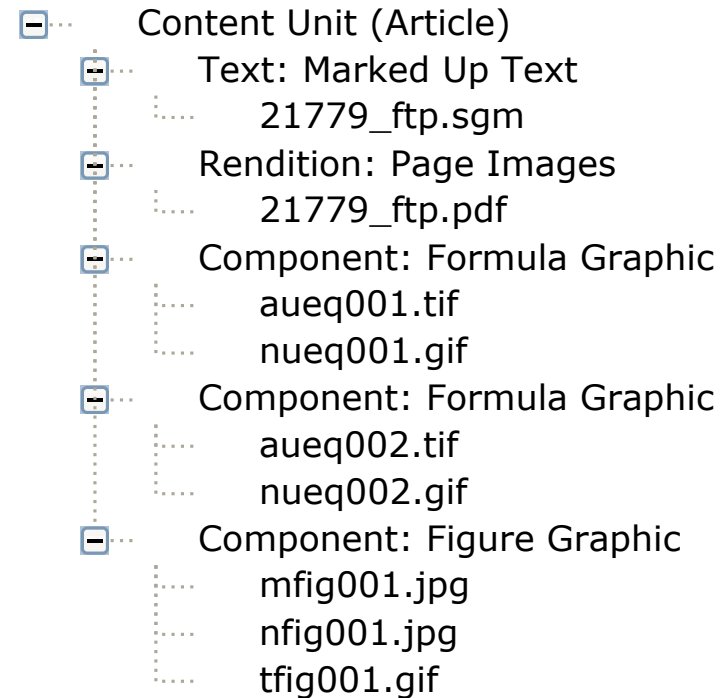


# Portico XML Configuration: Packages & Profiles

## Incoming File System



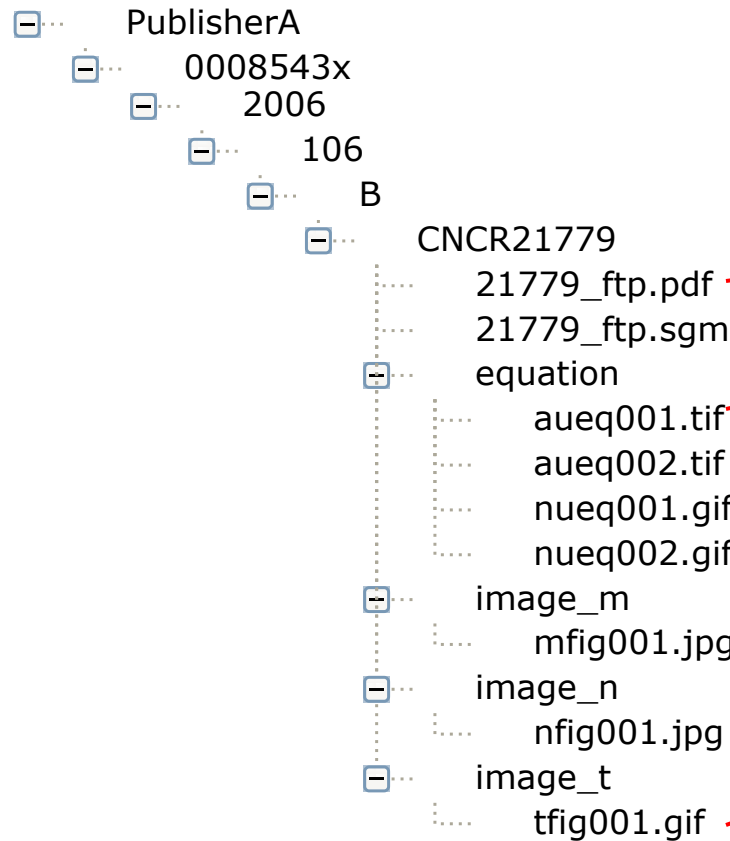
## Resulting Content Model



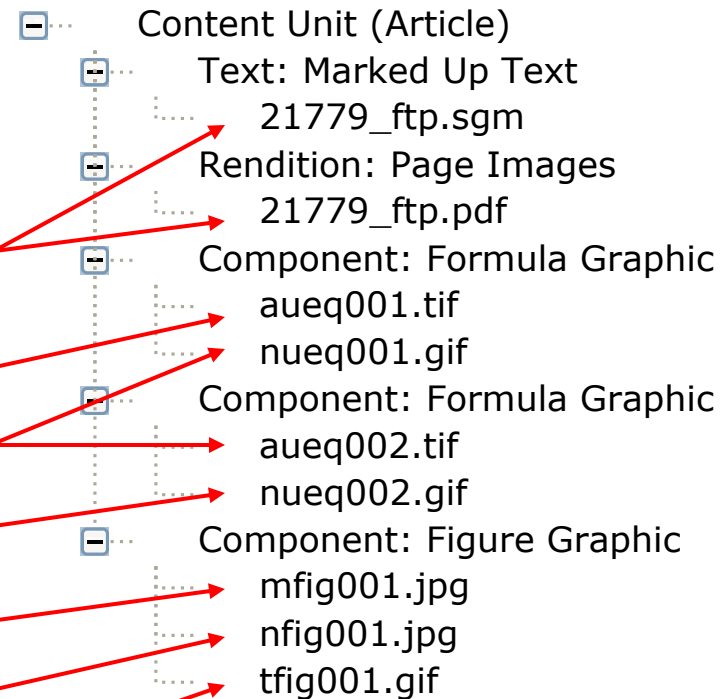


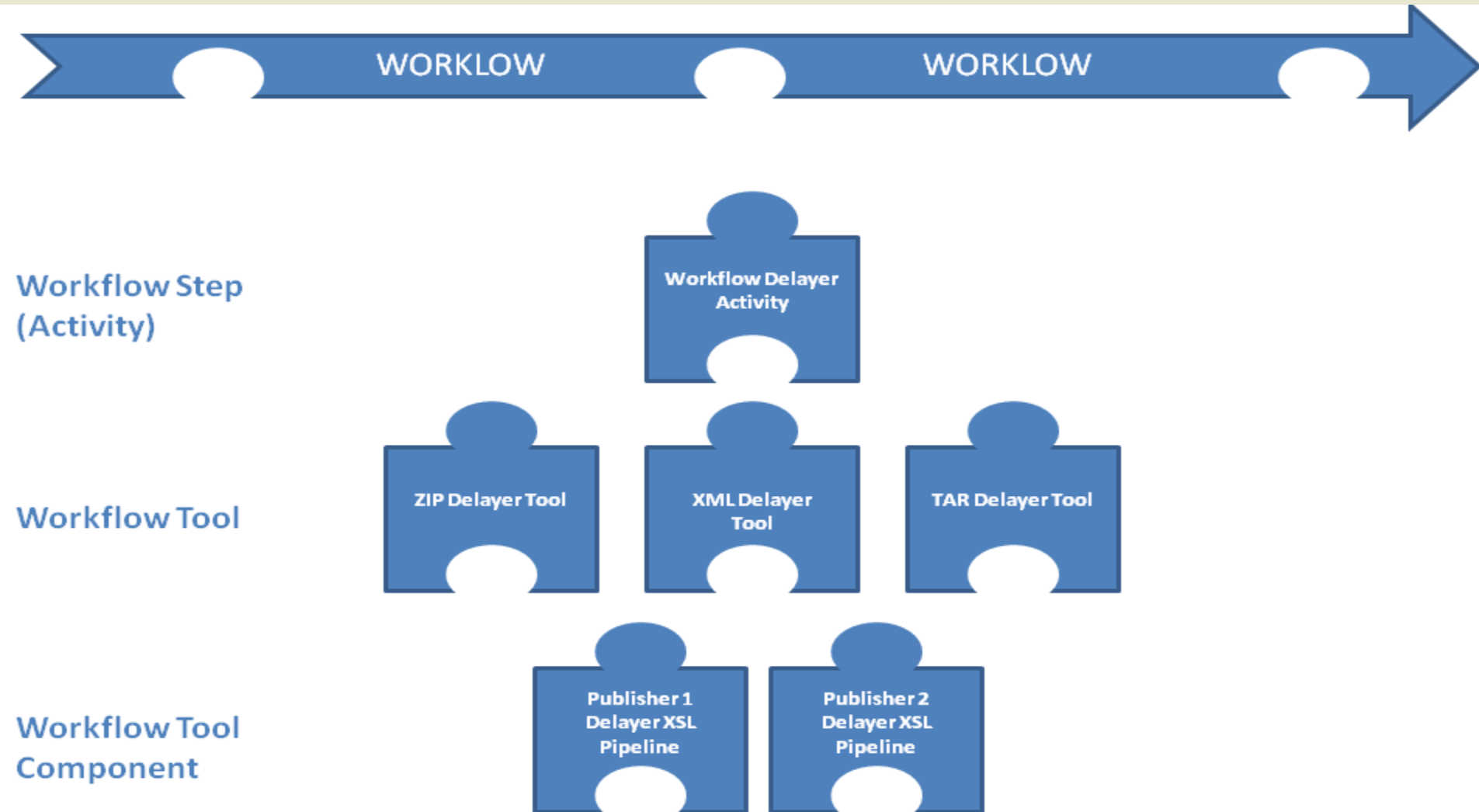
# Portico XML Configuration: Packages & Profiles

## Incoming File System

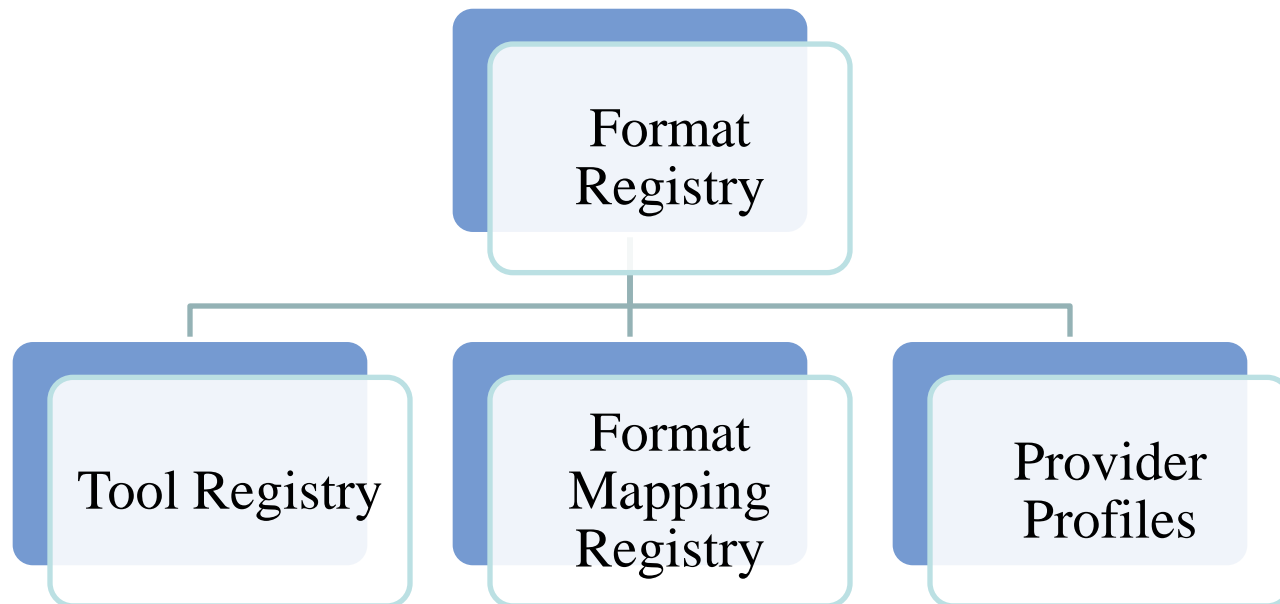


## Resulting Content Model





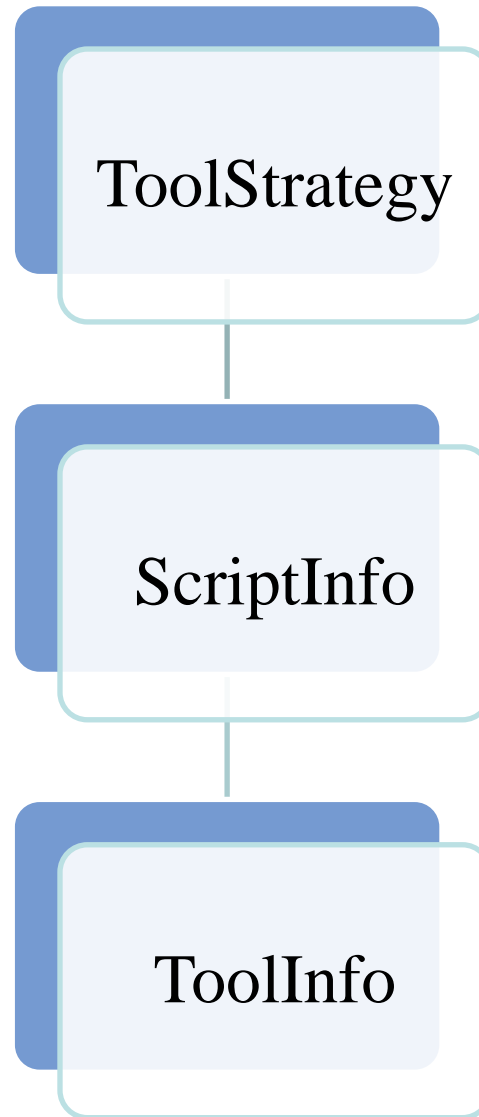
# Format-Driven Workflow: Format Registry



```
<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>
```

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

# Format-Driven Workflow: Tool Registry



# Format-Driven Workflow: Tool Registry

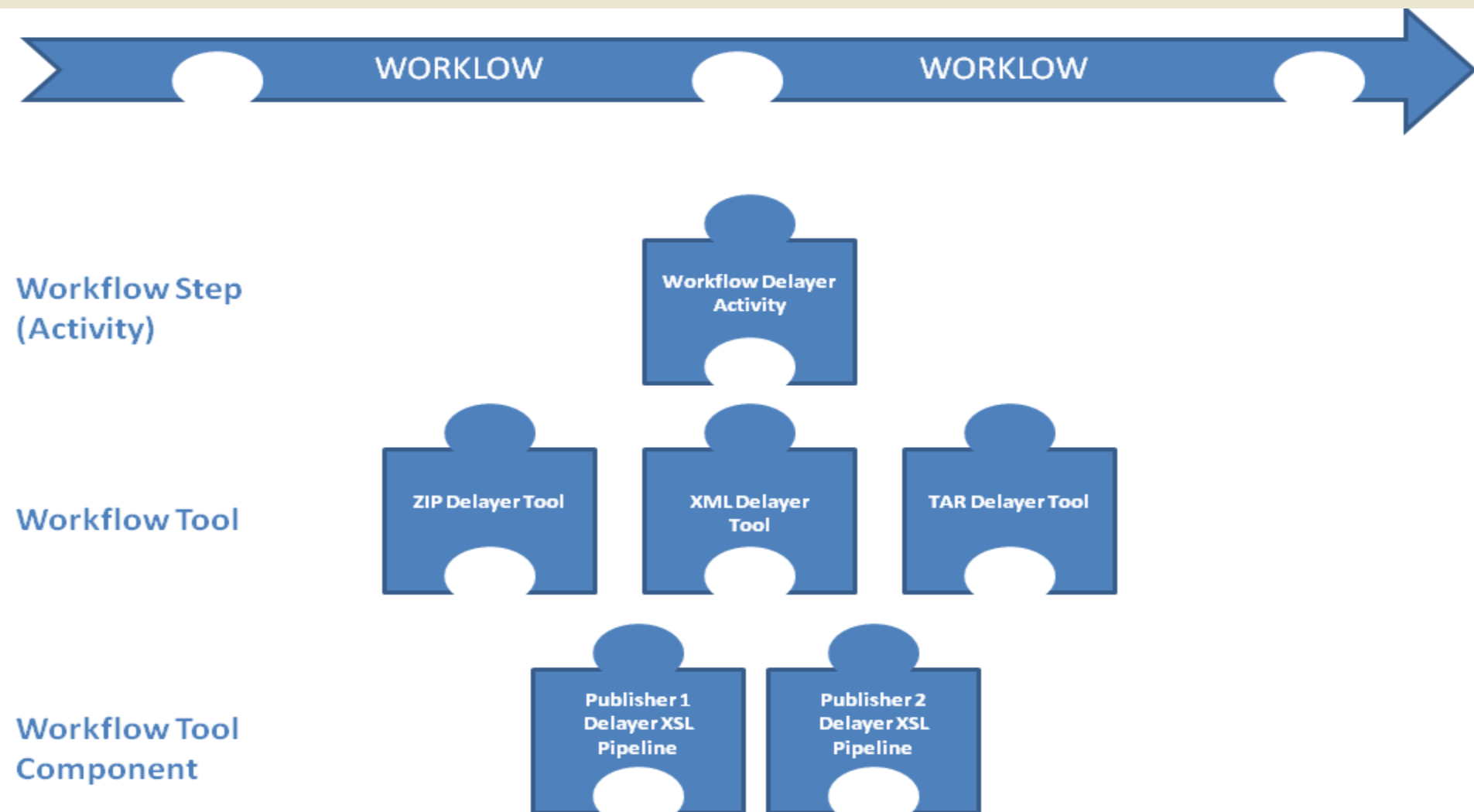
```
<ScriptInfo ScriptId="scrxxx">
  <Tool Rid="BaseTransform_1.0">
    <Parameters>
      <Parameter>
        <Name>StyleSheetList</Name>
        <ValueOrderedList>
          <ValueOrderedListItem>
            <Number>10</Number>
            <Value>xxx2ptc_1_10.xsl</Value>
          </ValueOrderedListItem>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



```
<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> ...
```





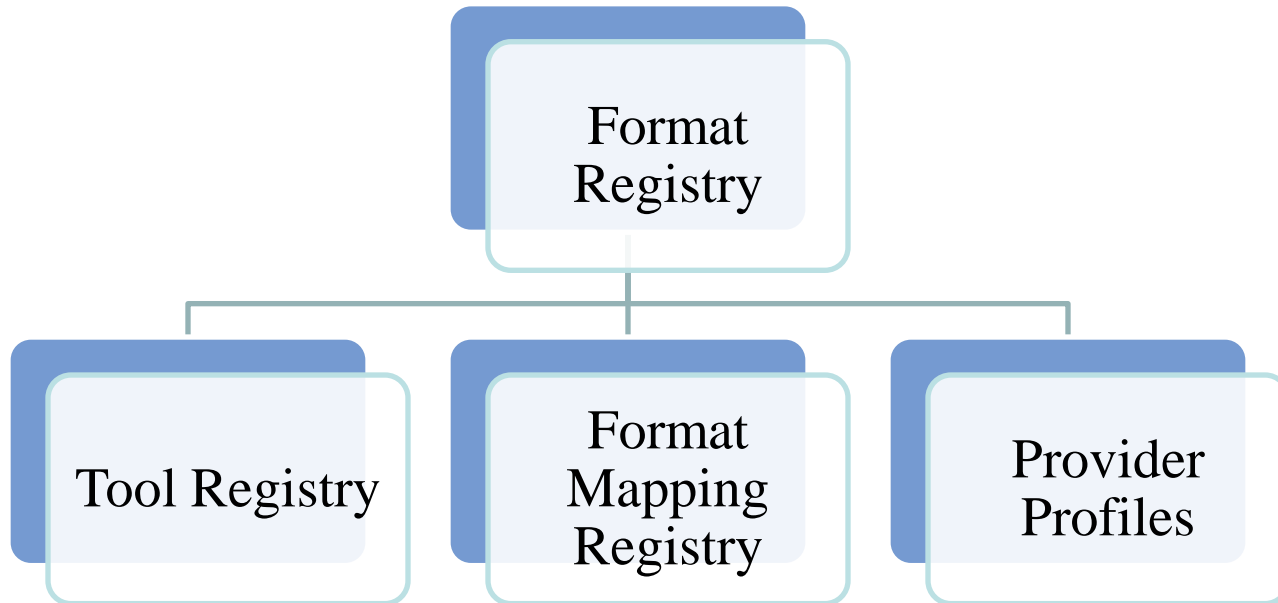


## ❖ Consistency

```
<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("");
        }
    }
}
```

## ❖ “Referential Integrity”



```

ev/workarea/shared
├── Tools
│   ├── CVS
│   │   ├── src
│   │   │   ├── CVS
│   │   │   │   ├── org
│   │   │   │   │   ├── CVS
│   │   │   │   │   │   ├── portico
│   │   │   │   │   │   │   ├── CVS
│   │   │   │   │   │   │   │   ├── tool
│   │   │   │   │   │   │   │   │   ├── CVS
│   │   │   │   │   │   │   │   │   ├── data
│   │   │   │   │   │   │   │   │   ├── deploy
│   │   │   │   │   │   │   │   │   └── scripts
│   │   │   │   │   │   │   │   │   ├── audmdextraction
│   │   │   │   │   │   │   │   │   ├── checksum
│   │   │   │   │   │   │   │   │   ├── CVS
│   │   │   │   │   │   │   │   │   ├── datasetextraction
│   │   │   │   │   │   │   │   │   ├── delayering
│   │   │   │   │   │   │   │   │   ├── delayeringkluer
│   │   │   │   │   │   │   │   │   ├── dmdcuration
│   │   │   │   │   │   │   │   │   ├── dmdextraction
│   │   │   │   │   │   │   │   │   ├── filereferenceextraction
│   │   │   │   │   │   │   │   │   ├── filereferencereplacement
│   │   │   │   │   │   │   │   │   ├── htmlrendition
│   │   │   │   │   │   │   │   │   ├── identification
│   │   │   │   │   │   │   │   │   ├── jhove_1_2
│   │   │   │   │   │   │   │   │   ├── qcddetection
│   │   │   │   │   │   │   │   │   ├── tmdextraction
│   │   │   │   │   │   │   │   │   ├── transform
│   │   │   │   │   │   │   │   │   ├── transformation
│   │   │   │   │   │   │   │   │   ├── validation
│   │   │   │   │   │   │   │   │   ├── viruscheck
│   │   │   │   │   │   │   │   │   ├── xfileextraction
│   │   │   │   │   │   │   │   │   ├── xfilequalitycheck
│   │   │   │   │   │   │   │   │   └── xfilequalitychecker

```

A screenshot of a Windows File Explorer window displaying a folder named "shared". The folder contains a large number of files, each represented by a yellow folder icon. The files are listed in alphabetical order and include various document types such as DTDs (Document Type Definitions), XML files, and HTML files. Notable files include ".svn", "AAA\_NLM\_Journal\_Archiving\_DTD\_1.0", "AACCA\_AIP\_SPIN\_Article\_DTD\_2.1", "AAS\_IOP\_Reference\_List\_DTD\_1.0", "ACM\_Proceeding\_XML\_DTD\_5.0", "ACS\_Attypon\_ACHS\_NLM\_Journal\_Archiving\_DTD\_2.2", "AGU\_Article\_XML\_DTD\_3.2", "AIA\_Attypon\_NLM\_Journal\_Archiving\_DTD\_1.0", "AIP\_Article\_DTD\_1.5", "AIP\_Article\_DTD\_2.0", "AIP\_Article\_DTD\_3.0", "AIP\_Article\_DTD\_3.1.5", "AIP\_Article\_DTD\_3.2", "AIP\_Article\_DTD\_3.22", "AIP\_Article\_DTD\_3.23", and "AIP\_Book\_XML\_DTD\_1.0". The background of the window shows a blurred view of a desk with a computer monitor and keyboard.

The screenshot shows a Java IDE with a project structure on the left and a list of JAR files in the center. The project structure includes a 'setup' folder containing 'bin', 'config', and 'lib' subfolders. The JAR list contains the following files:

- asm.jar
- asm-attrs.jar
- c3p0-0.9.1.jar
- cglib-2.1.3.jar
- com.noelios.restlet.jar
- commons-beanutils.jar
- commons-collections-3.2.1.jar
- commons-configuration-1.6.jar
- commons-io-1.4.jar
- commons-lang-2.4.jar
- commons-logging.jar
- commons-net-1.4.1.jar
- ConpreprepHibernateConfig.jar
- deepExportService.jar
- dom.jar
- dom4j-1.6.1.jar
- ehcache-1.2.3.jar
- FastInfoset.jar
- hibernate3.jar
- HibernateConfig.jar
- imap.jar
- imq.jar
- jaas.jar
- jacksum.jar
- jaxb1-impl-2.0.jar
- jaxb-api-2.2.jar
- jaxb-impl-2.2.jar

```
<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>
```

## Choices

- Properties files
- Relational DB



## Limits

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

## Aesthetics



Image: [Wikimedia Commons](http://commons.wikimedia.org/wiki/File:River_terrapin.jpg)  
([http://commons.wikimedia.org/wiki/File:River\\_terrapin.jpg](http://commons.wikimedia.org/wiki/File:River_terrapin.jpg))



PORTICO

## Questions or Comments?

Sheila Morrissey  
[sheila.morrissey@ithaka.org](mailto:sheila.morrissey@ithaka.org)  
[@sheilaMorr](https://twitter.com/sheilaMorr)  
[www.portico.org](http://www.portico.org)



PORTICO