Automated Workflow for the Ingest and Preservation of Electronic Journals

Evan Owens
Chief Technology Officer
Portico
evan.owens@portico.org

Portico Background

- Trusted third party archiving solution
- Funding from libraries, publishers, foundations, etc.
- Initial pilot described at Archiving 2004
- Business and access model heavily revised
- Launched to publishers in September 2005
- Launched to libraries in January 2005

- 13 publishers signed; more to follow soon
- ~3500 titles; ~7M articles
- See www.portico.org for latest information

Electronic Journals and Digital Preservation

- Journal publishing models are evolving
  - Publishing practice varies:
    - Print only, E-only, both
    - More / less / same in each edition
  - E-product varies:
    - HTML Header & PDF
    - HTML Full-text with links and supplemental stuff & PDF
    - HTML only
- A “work” with multiple “manifestations”
  - XML or SGML source files
  - Print PDF used to drive printing press
  - Web PDF optimized for online delivery
  - HTML header or full text
    - Often generated from XML or SGML source

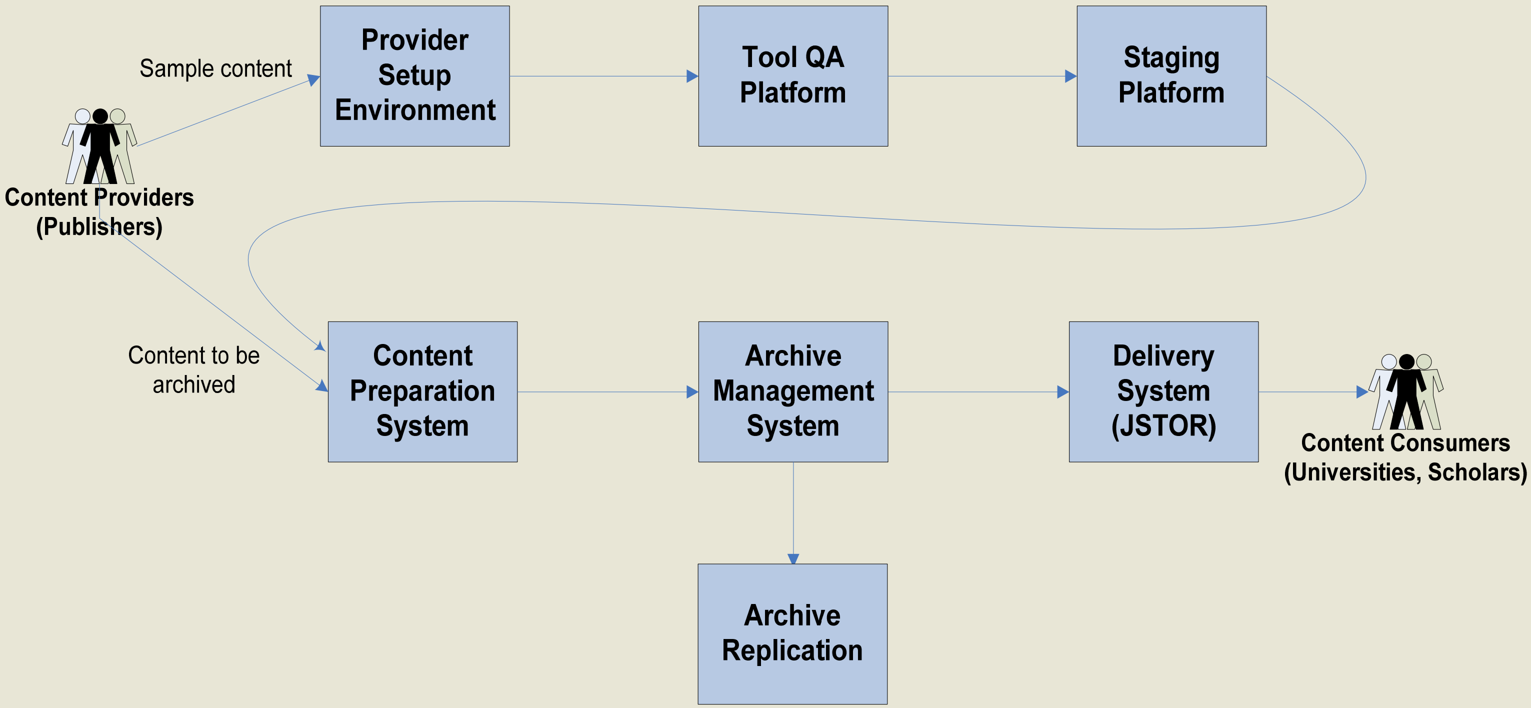
Portico Archival Strategy for E-Journals

- Source file archiving
  - Preserve the components not the rendition
  - Include high-resolution files (PDF and figures) if available
  - All e-only components (data, media, etc.)
  - SGML / XML structured text by preference
    - HTML as last resort
- Preserve intellectual content not “look and feel” of HTML
  - HTML renditions are an artifact of current (and past) technology
    - Often dynamically generated
    - Fragile technology, overdue for change
- Preserve only essential features of the user interface
  - Reference linking, other content-based features
  - Not generic navigation or search or e-commerce features

Electronic Journal Data Issues

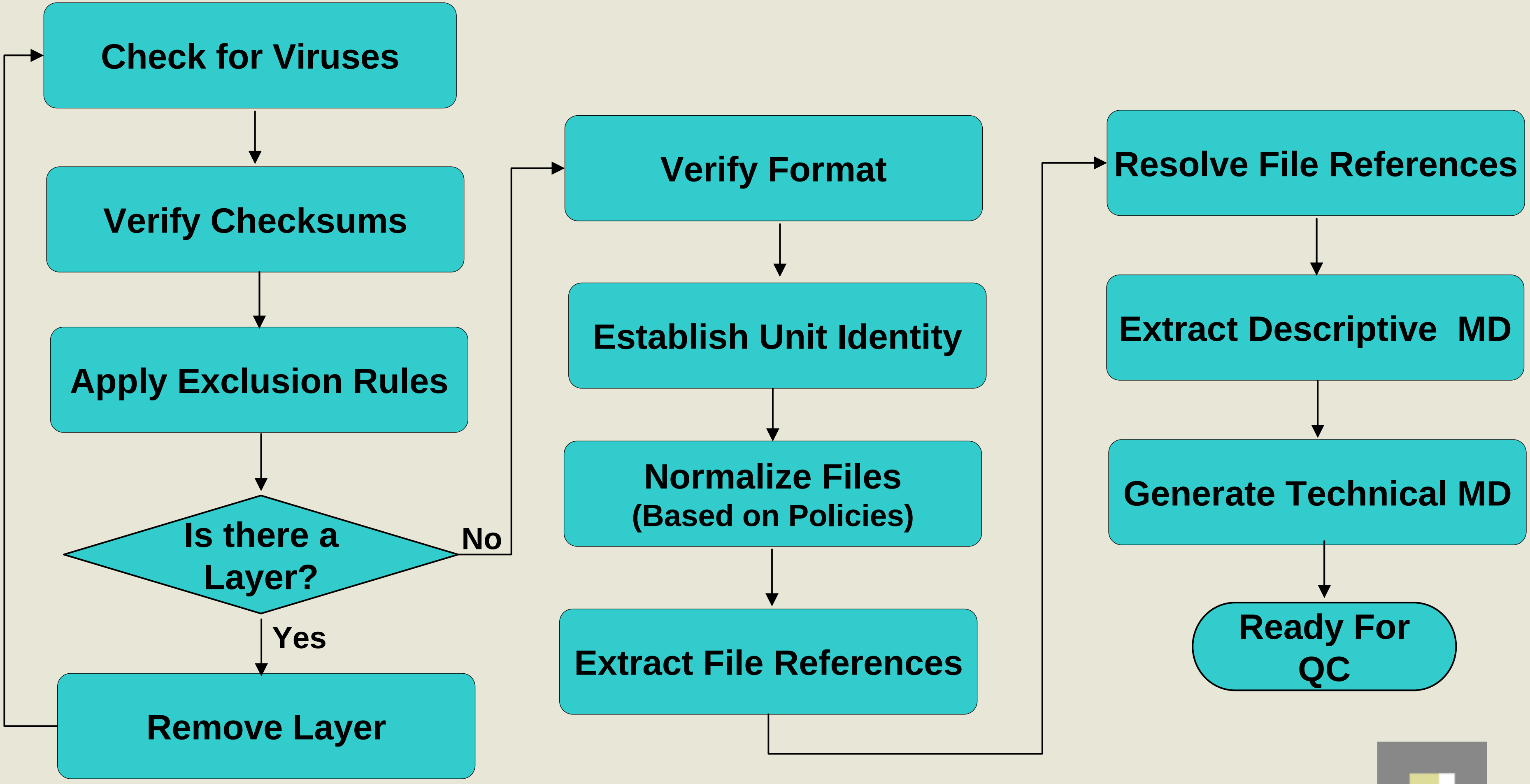
- Inputs
  - Whatever the publisher wants to send us
  - However the publisher wants to send it to us
  - Per article: one text or metadata file, zero or more other files
  - Arbitrary (publisher-specific) collections of data
    - Proprietary file & directory naming conventions
    - Proprietary formats
  - Undocumented business rules hidden in the data
- Output to Archive
  - Normalized content (pre-emptive migration of proprietary formats)
  - Automated metadata capture/generation: technical, descriptive, events
  - Packaged in Portico's variant of METS

Process Overview


Sample content
Provider Setup Environment
Tool QA Platform
Staging Platform
Content Providers (Publishers)
Content to be archived
Content Preparation System
Archive Management System
Delivery System (JSTOR)
Content Consumers (Universities, Scholars)
Archive Replication

Automated Processing for E-Journal Content

(high-level summary)



Incoming File System

```
PublisherA
 └───0008543X
        └───2006
               └───106
                      └───8
                             └───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
```

EXCERPTS FROM SGML TEXT:

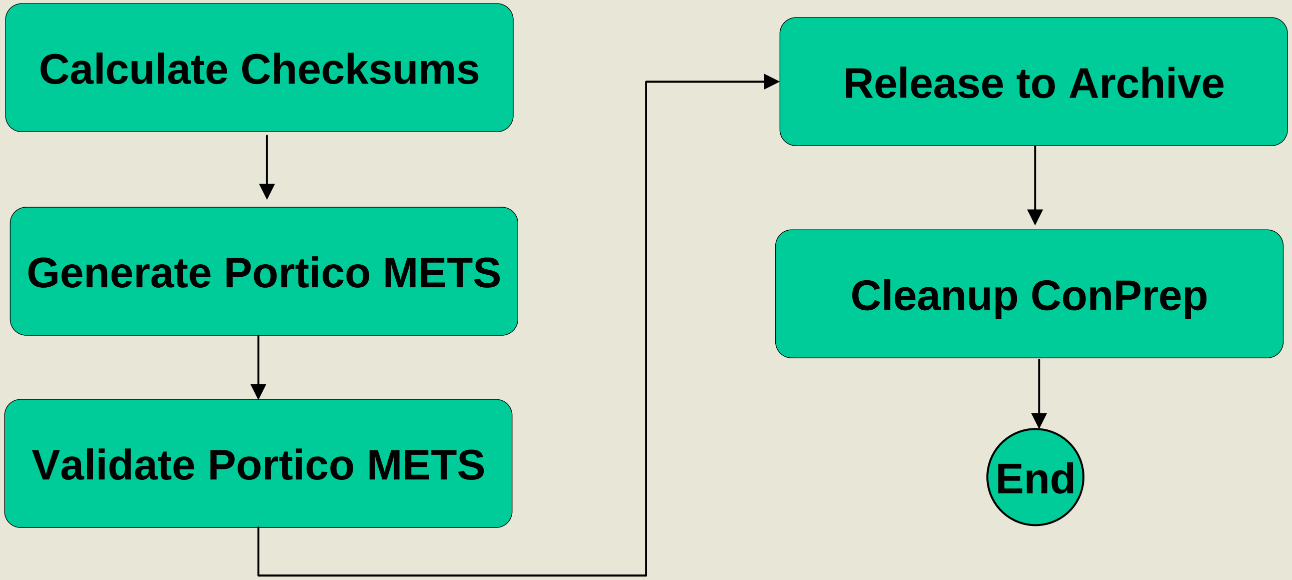
```
the following statistical model was fitted to the data: <UEQN NAME="ueq002" LOC="FIXED"></UEQN> in which <I>T</I> &equals; 1 if Grade 3 or 4 neutropenia was present

The overall survival for all patients is illustrated in Figure <FIGR HREF="fig1">1</FIGR>.

<FIG ID="fig1" LOC="FLOAT"><GRAPHIC NAME="fig001"></GRAPHIC><NUMBER>1</NUMBER><CAPTION><P>Overall survival for 160 eligible and evaluable patients with recurrent solid tumors who were enrolled on Children&apos;s Cancer Group Study 0962.</P></CAPTION></FIG>
```

EXCERPTS FROM NORMALIZED XML TEXT:

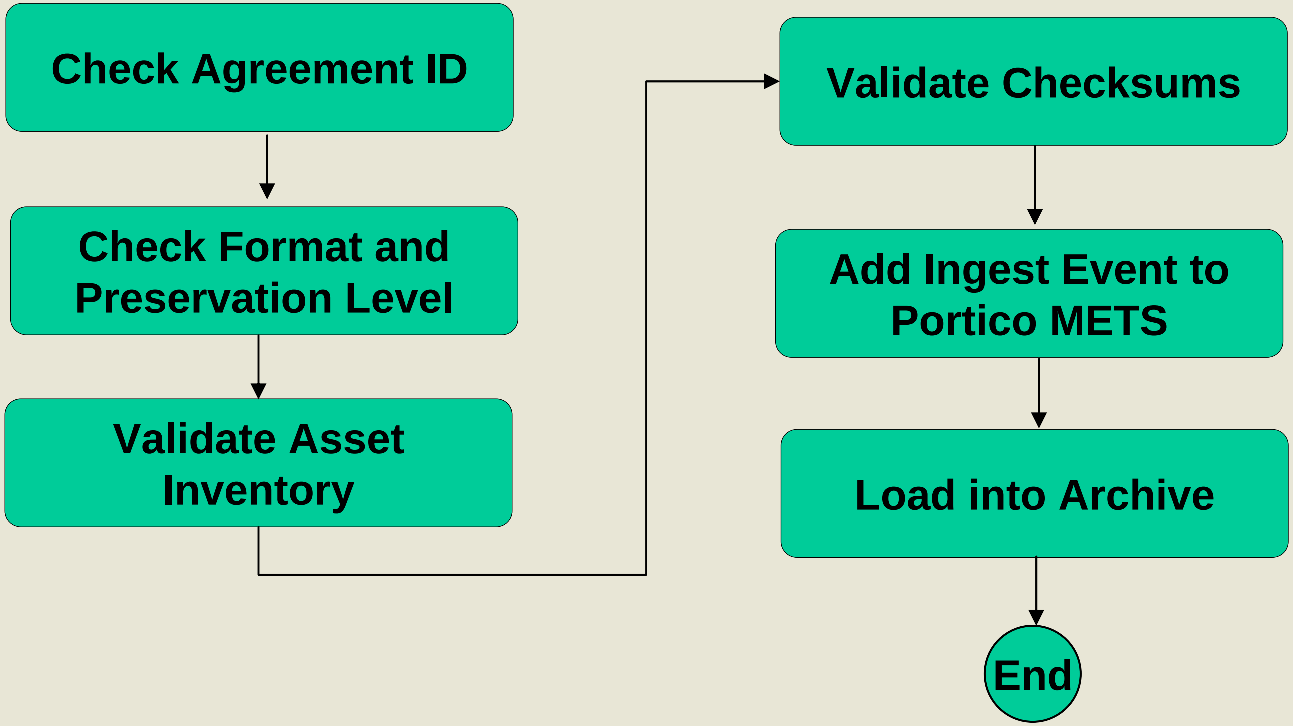
```
the following statistical model was fitted to the data: <disp-
formula><graphic xlink:href="ark:/27927/pc01mtb5t"
position="anchor"/></disp-formula>in which <italic>T</italic> = 1
if Grade 3 or 4 neutropenia was present

The overall survival for all patients is illustrated in Figure
<xref rid="FIG1" ref-type="fig">1</xref>.</p>

<fig fig-type="figure" id="FIG1" position="float"><label>
<x x-type="archive">Figure</x>1<label><caption><p>Overall survival
for 160 eligible and evaluable patients with recurrent solid tumors
who were enrolled on Children's Cancer Group Study 0962.</p></
caption><graphic position="anchor"
xlink:href="ark:/27927/pc01mtbqj"></graphic></fig>
```

Automated Processing after QC

(for all content types)



Archive Ingest Processing



Check Agreement ID

Check Format and Preservation Level

Validate Asset Inventory

Validate Checksums

Add Ingest Event to Portico METS

Load into Archive

End

Some Critical Issues

- Content isn't perfect
  - Must have policies and workflow for invalid data
  - There are degrees of "badness"
  - Strict format validity does not equate to usefulness or usability
    - E.g., Well-formed but not valid PDF
    - E.g., Valid PDF with bad embedded font
    - E.g., Invalid JPEG
- Content creation practices change over time
  - Publishers (content providers) aren't consistent
  - Or don't warn you that they are changing something
  - Defensive programming required
- Software isn't perfect
  - Assume that there will be internal failures
  - Reversibility and audit trail are essential
    - Portico Tool Registry and events metadata

A Sample Tool Event

```
<EventTransformedFile Timestamp="2006-05-22T11:39:46.830-04:00">
  <Tool>
    <ToolInfo>
      <RunDate>2006-05-22T11:39:46.150-04:00</RunDate>
      <ToolWrapper>
        <RegisteredName>BepressTransformTool:1.0:2006-04-21</RegisteredName>
        <RuntimeEnv>Java:Sun Microsystems
Inc.:1.5.0_04:SunOS:sparc:5.9</RuntimeEnv>
        <DependentLibSet>
          <DependentLib>bepress.xsl</DependentLib>
          <DependentLib>insert-titles.xsl</DependentLib>
          <DependentLib>insert-portico-doctype.xsl</DependentLib>
          <DependentLib>nlmpub2_1_to_ptc2_0.xsl</DependentLib>
          <DependentLib>porticoCommon_1_1.xsl</DependentLib>
          <DependentLib>gentext.xsl</DependentLib>
        </DependentLibSet>
      </ToolWrapper>
      <VendorTool>
        <VendorToolName>com.icl.saxon.TransformerFactoryImpl</VendorToolName>
        <RuntimeEnv>Java:Sun Microsystems
Inc.:1.5.0_04:SunOS:sparc:5.9</RuntimeEnv>
      </VendorTool>
    </ToolInfo>
  </Tool>
</EventTransformedFile>
```

Envoi

- Digital preservation as “interoperability with the future”
  - Let’s test now while we can still recover
  - If we can’t move content from party to party today,
  - Why do we think we will be able to in the future?
- Data exchange is valuable
  - To both parties
  - There is knowledge transfer as well as data transfer
  - Standards are only standard when practiced, in use
- Robust electronic content production and content management systems will help to make preservation easier and cheaper
  - We have to get ahead of the problem, not just clean up the mess afterwards