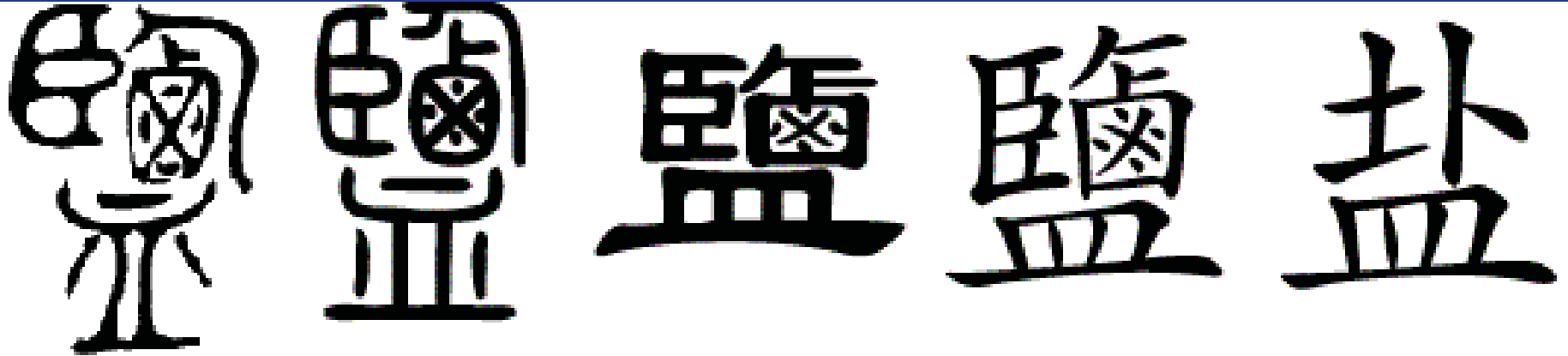


Archivists' Workbench: A Framework for Testing Preservation Infrastructure

Richard Marciano

Sustainable Archives & Library Technologies (SALT) Lab



San Diego Supercomputer Center (SDSC)

University of California San Diego (UCSD)

marciano@sdsc.edu

Relating InterPARES Research and an AW Framework

- Policy Analysis
- Description
- Terminology

- Modeling
 - Functional models
 - Data flow models

- Digital infrastructure

“Antarctic Treaty Searchable Database Case Study”

Paul Berkman (UCSB)

→ What is the appropriate level of granularity to discover meaningful relationships in the digital collection?

→ What is the impact of the discovery on the policies themselves

Persistent Archives Testbed (PAT)

- Test a *community model* for electronic records management, with archival and technological functions in a distributed network (data grid technology)
- The processes that will be automated are:
 - appraisal,
 - accessioning,
 - arrangement,
 - description,
 - preservation &
 - access.

Goal

- **Initial test sites:**

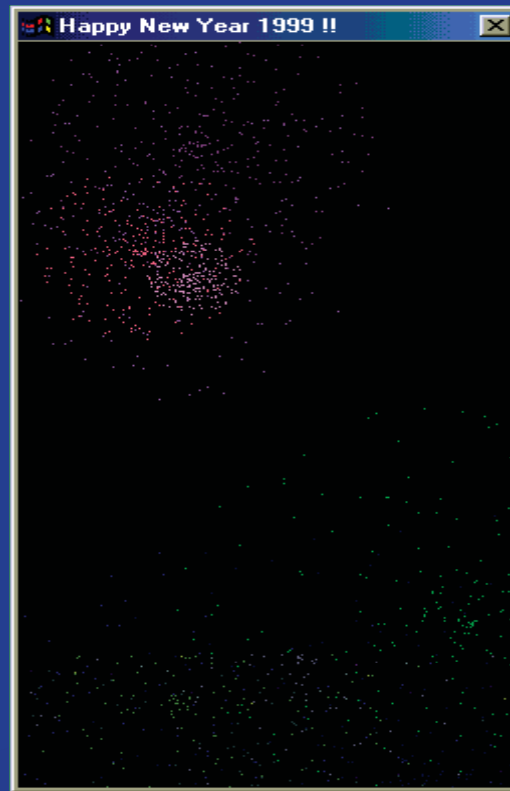
- (1) Michigan Department of History, Arts and Libraries,
- (2) Ohio Historical Society,
- (3) Kentucky Department for Libraries and Archives,
- (4) Minnesota Historical Society,
- (5) Stanford Linear Accelerator Archives and History Office.

- **Additional partners:**

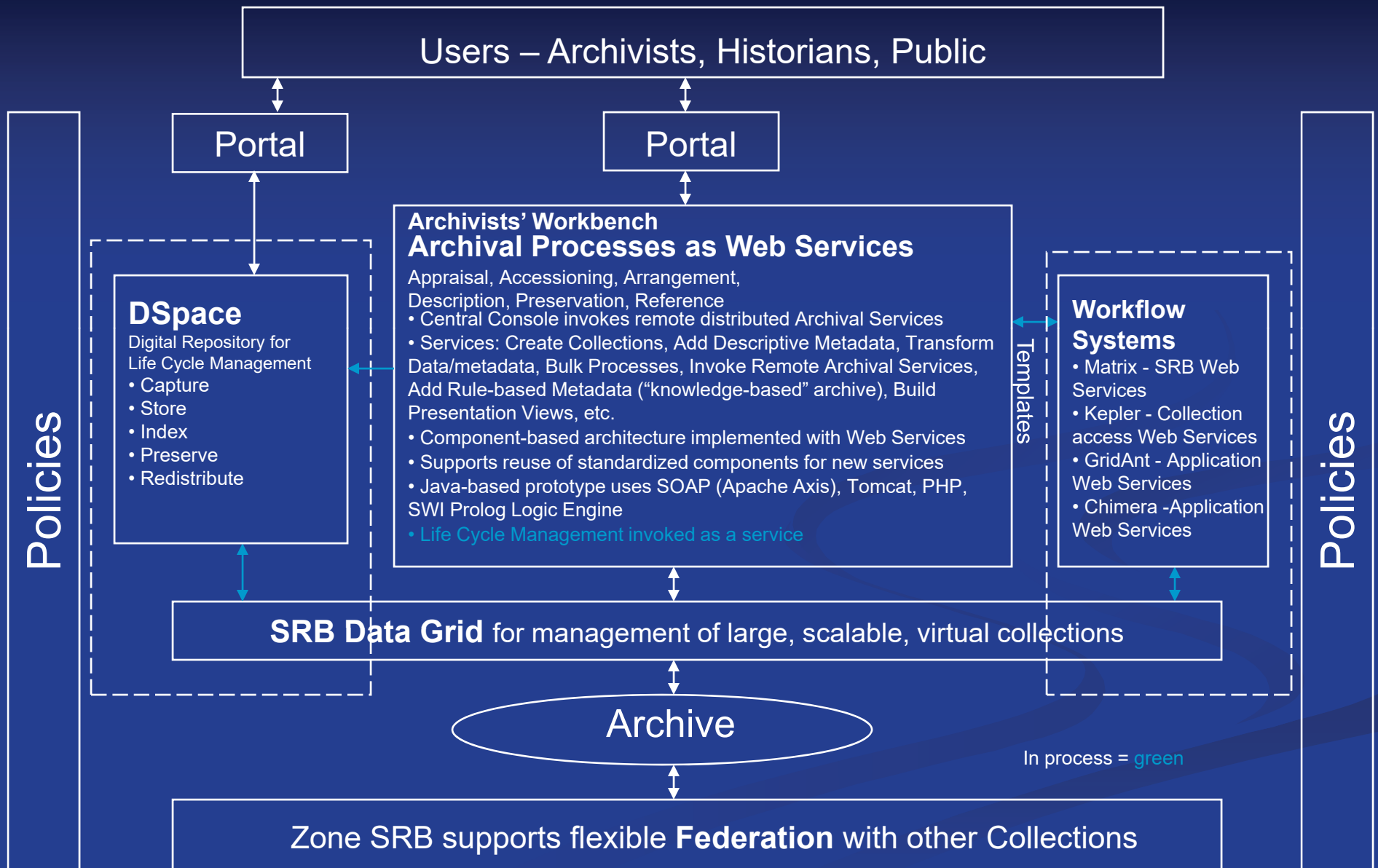
- 1) Yale Manuscript Archives
- 2) University of Illinois at Urbana-Champaign
- 3) Kansas Historical Society
- 4) UCLA - CIE

Ohio OBES e-mail Collection

... an example of issues related to POLICY
item-level vs. collection-level appraisal



SDSC Prototype Archivists' Workbench



Framework Components

Archivists' Workbench
Archival Processes as Web Services

Portal Technology

Workflow Systems

Data Grids & Federation

Batch1

Shipment #1 contains 1,241 files (a.k.a. documents) comprising the following file types:

File Count	File Type
602	TMP
416	<u>doc / DOC</u>
92	<u>txt</u>
48	<u>xls</u>
23	822
19	<u>htm</u>
13	<u>pdf / PDF</u>
9	<u>ppt</u>
4	<u>jpg</u>
4	<u>mdb</u>
4	<u>dot</u>
2	<u>rtf</u>
2	<u>dat</u>
2	<u>bmp</u>
1	<u>opx</u>
TOTAL: 1,241 files	TOTAL: 15 types

Batch2

Shipment #2 contains 100 files comprising the following file types:

File Count	File Type
53	TXT / <u>txt</u>
31	<u>doc / DOC</u>
5	822
3	<u>htm</u>
2	KEY
1	<u>pdf</u>
1	000
1	<u>gif</u>
1	<u>jpg</u>
1	098
1	<u>mem</u>
TOTAL: 100 files	TOTAL: 11 types

Appendix A: Preservation System Functional Requirements (1-25)

1. The system shall provide the capability to preserve the archival record for the life of the record.
2. The system shall provide the capability to re-present the original content, context and structure of the archival record.
3. The system shall provide the capability to make the archival record available for printing, viewing and saving.
4. The system shall provide the capability to store the archival record with all of its attributes.
5. The system shall provide the capability to associate additional attributes (a.k.a. "preservation attributes") with the archival record.
6. The system shall provide the capability to associate attributes of an archival record with preservation attributes.
7. The system shall provide the capability to populate preservation attributes with data from the archival record and its received attributes.
8. The system shall provide the capability to place "zero," "null" or other authorized user value in a preservation attribute when archival record attribute value is invalid (e.g. not present – null or zero value itself, of incorrect type – letters instead of numbers, exceeds preservation attribute field length).
9. The system shall provide the capability to output for printing, viewing and saving a "preservation attribute to archival record attribute" outcome report for each archival record accessioned.
10. The system shall provide the capability to make available for printing, viewing and saving all attributes of the archival record.
11. The system shall provide the capability to populate received attributes of a non-electronic archival record managed by an RMA.
12. The system shall provide the capability to populate received attributes of the archival record.
13. The system shall maintain the authenticity of the record for as long as is it retained.
14. The system shall provide the capability to output for printing, viewing and saving a copy of the record and all its components (e.g. attachments).
15. The system shall provide the capability to manage each record with a unique identifier.
16. The system shall provide the capability to output for printing, viewing and saving all attribute data of the records.
17. The system shall provide the capability to output for printing, viewing and saving all attribute names of the records.
18. The system shall provide the capability to search on all attribute names, attribute values, index terms and data contained in the contents of the archival record.
19. The system shall ensure an archival record and its received attributes cannot be modified.
20. The system shall provide the capability to modify the preservation attributes on the Modifiable Preservation Attribute List.
21. The system shall provide the capability for an authorized user to modify an attribute in the Modifiable Preservation Attribute List.
22. The system shall provide the capability to output for printing, viewing and saving an audit of changes made to attributes in the Modifiable Preservation Attribute List.
23. The system shall provide the capability to output for printing, viewing and saving the record, all of its associated attributes, and attribute values.
24. The system shall provide the capability to output for printing, viewing and saving "search" attribute names (e.g. the title, name given to the data field) associated with the record.
25. The system shall provide access to the record based upon "access control rules".

XML Archiving & Packaging Tool (XAPT)

XAPT is a **Java-based** application that implements a central console mechanism. The architecture supports a suite of archival services and the implementation is based on Web Services technology.

The approach is compatible with recent developments in **“Grid” technology**, perceived by some as the the next evolution of the Web, where there is increasing emphasis on the network of resources and the “Web of Services” within which organizations work.

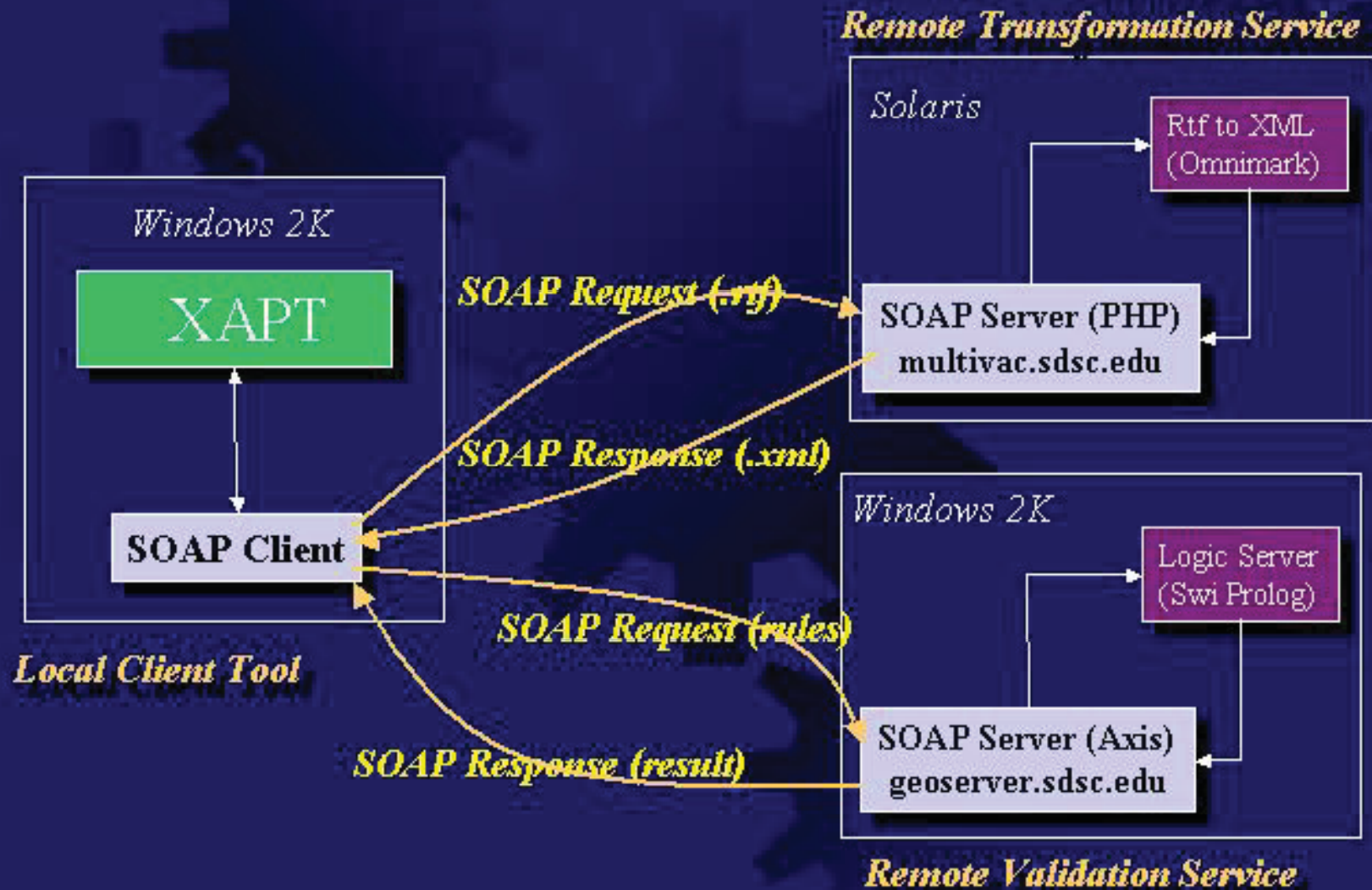
XAPT

- Borrows from **InterPARES** and an original idea from Bill Underwood on using JAR packages
 - ***“Preserving Authentic and Reliable Electronic Records in JARs”***, June 2000, a working paper by William E. Underwood, Georgia Institute of Technology, as part of the InterPARES Preservation Task Force. This paper explores the use of Java Archive files (JARs) as a mechanism to preserve electronic records.
 - Underwood, William E. ***“A Java JAR Implementation of an Archival Information Package,”*** Consultative Committee on Space Data Systems, XML Workshop, NASA Goddard, 20 August 2001.
- Based on **OAIS** model ideas
 - ***Open Archival Information System (OAIS) Reference Model***, <http://ssdoo.gsfc.nasa.gov/nost/isoas/>, January 2002. In the OAIS model, information packages are defined, including Archival Information Packages (AIPs).
- Defines an AIP or archival information package which contains a so-called KP or “Knowledge Package” made up of SEM + CON (SEMantics or logic rules / integrity constraints & CONtext or relationships to external information)
 - ***“Preservation of Digital Data with Self-Validating, Self-Instantiating Knowledge-Based Archives”***, B. Ludaescher, R. Marciano, R. Moore, ACM SIGMOD Record, 30(3), p. 54-63, 2001 (Special Issue on Advanced XML Data Processing), <http://www.sdsc.edu/~ludaesch/Paper/kba.pdf>

XAPT Basic Functionality

- The XAPT user should be able to:
 - create collections
 - add descriptive metadata
 - transform data/metadata
 - conduct bulk processing
 - Invoke remote archival services
 - add rule-based metadata (“knowledge-based” archive)
 - create Archival Information Packages (AIP) from collections
 - recreate collections from AIPs
- XAPT Architecture should be:
 - light-weight, portable, extensible, distributed, and service-oriented
- Archival Packages should be:
 - infrastructure independent/migration friendly

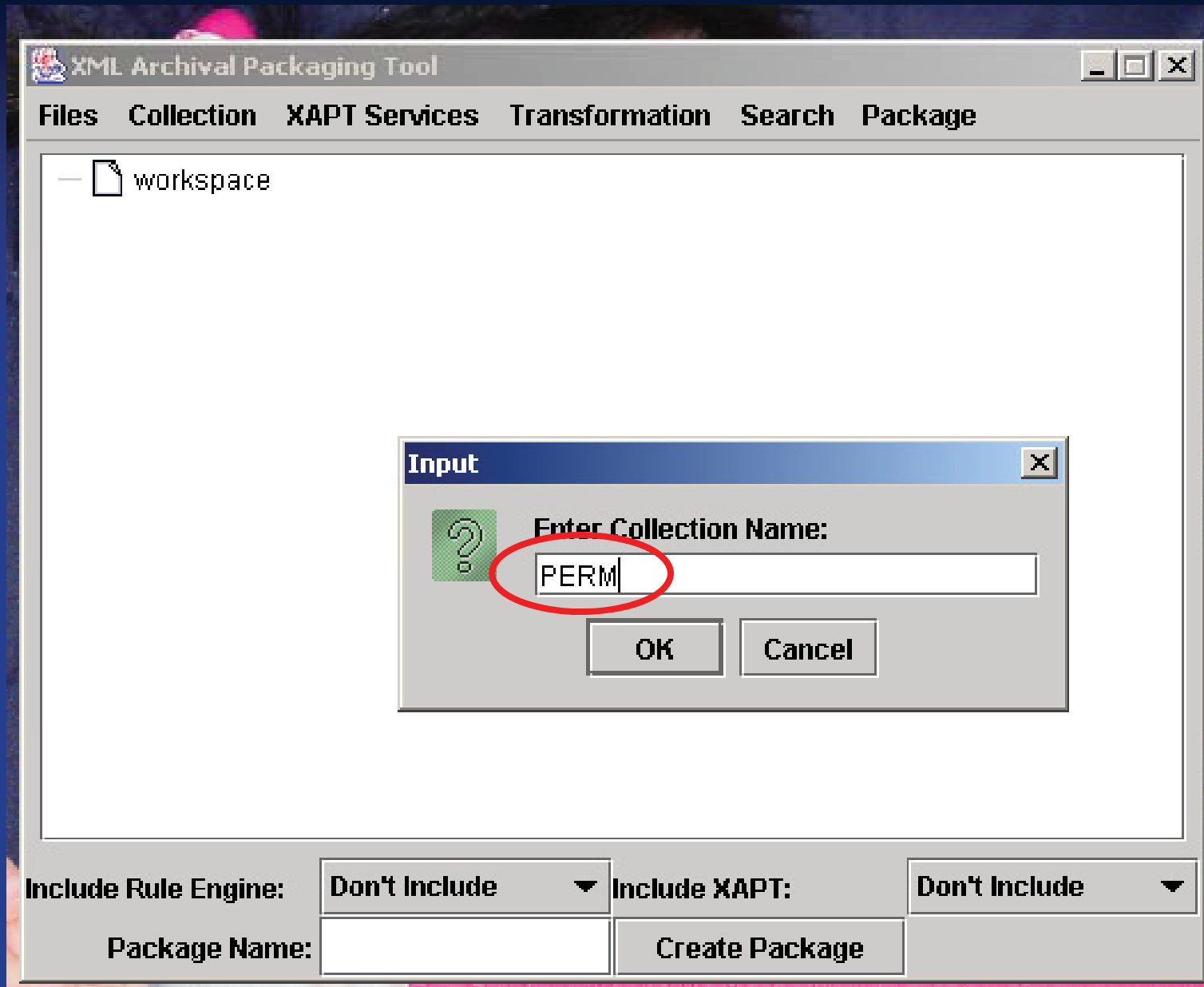
XAPT Distributed, Service-Oriented Architecture



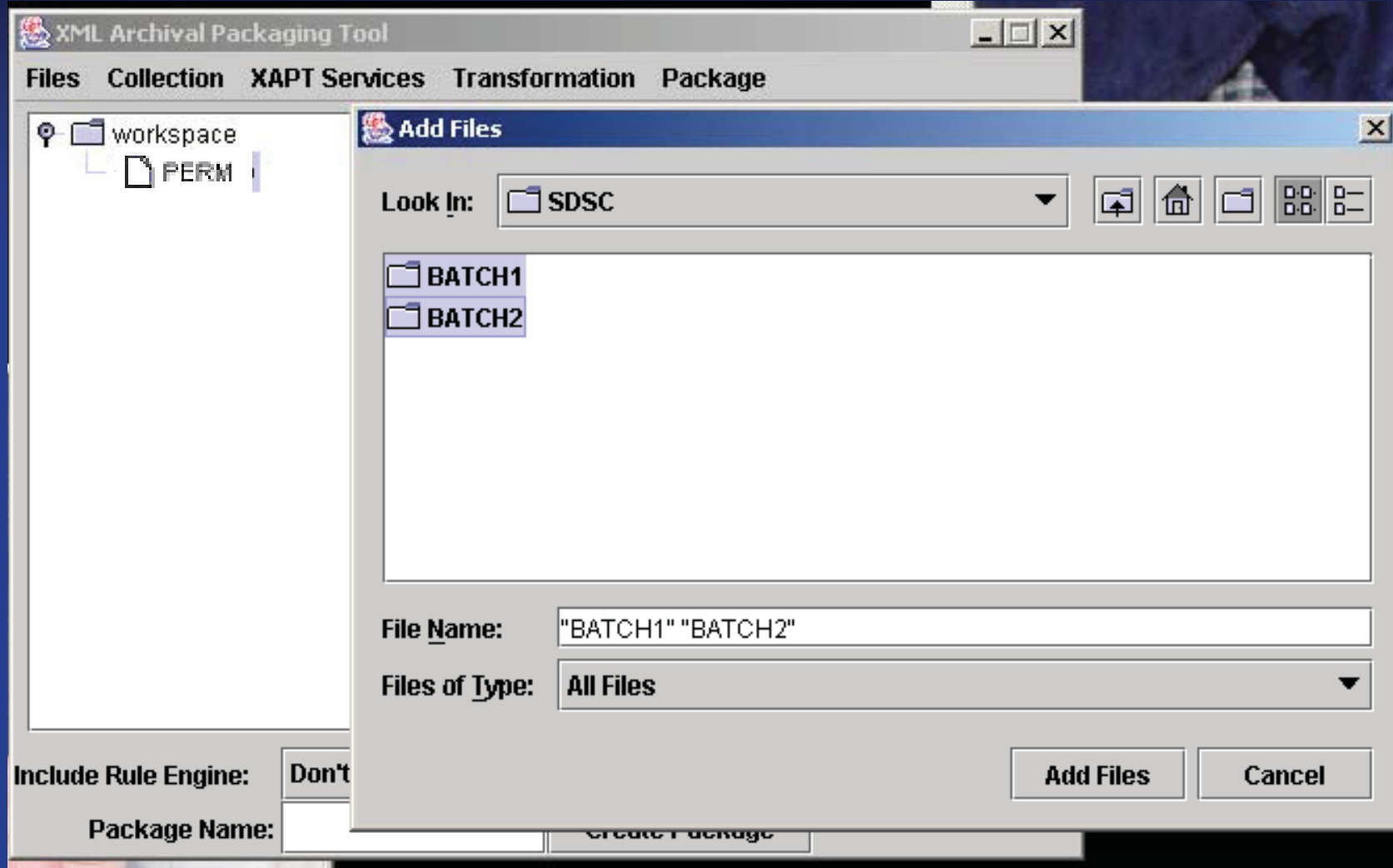
XAPT Walk-through

1. Create RMA Collection
2. Import RMA Records & Metadata
3. Create Collection Metadata
4. Transform RMA Metadata into Proposed PERM Standard
5. Perform Bulk Transformation of Email Records
6. Modify Preservation Metadata
7. Extract File Plan
8. Query the PERM Metadata
9. Create an RMA Archival Package
10. Reinstantiate the RMA Collection (“unpack”)

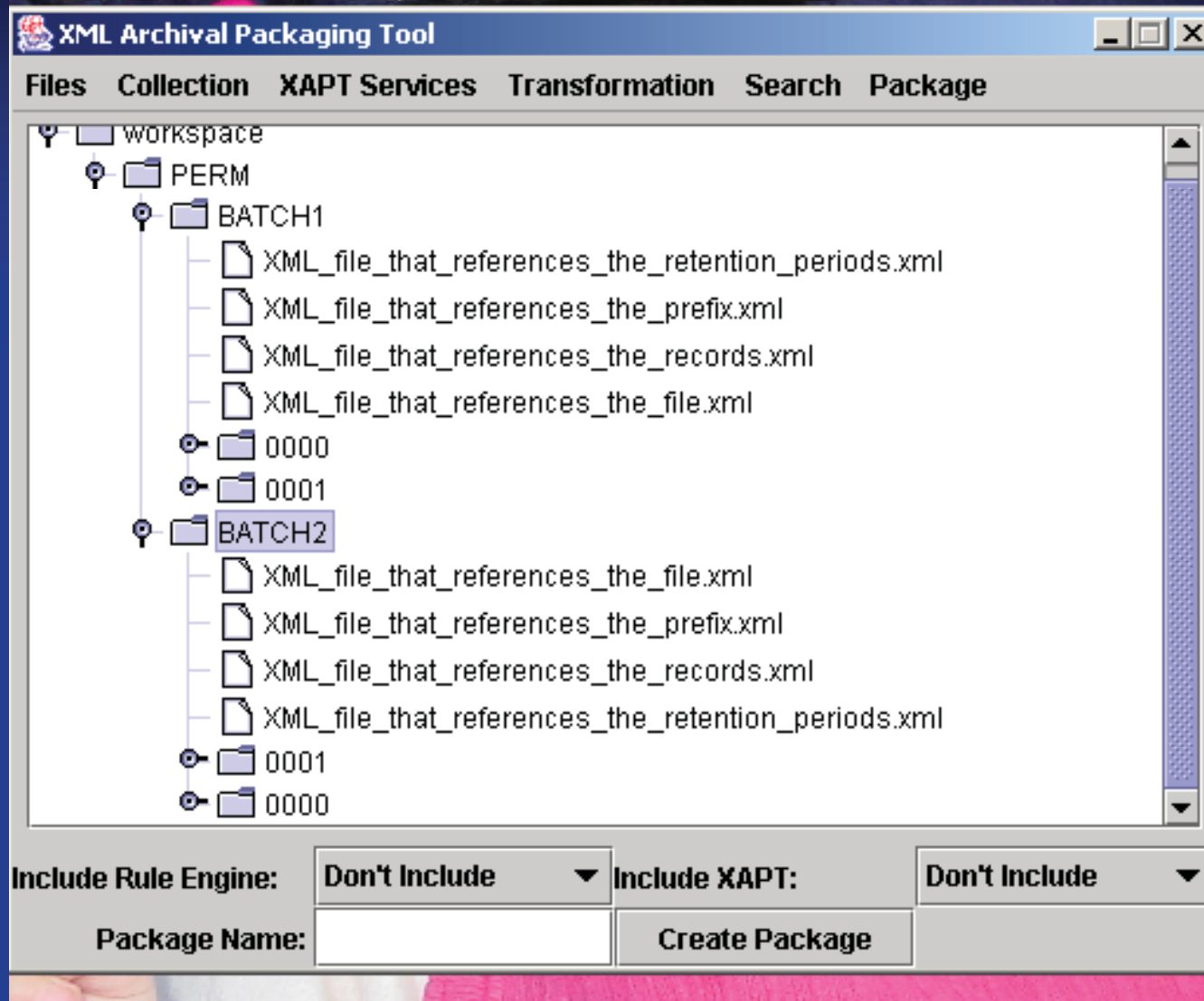
1. Create Collection PERM



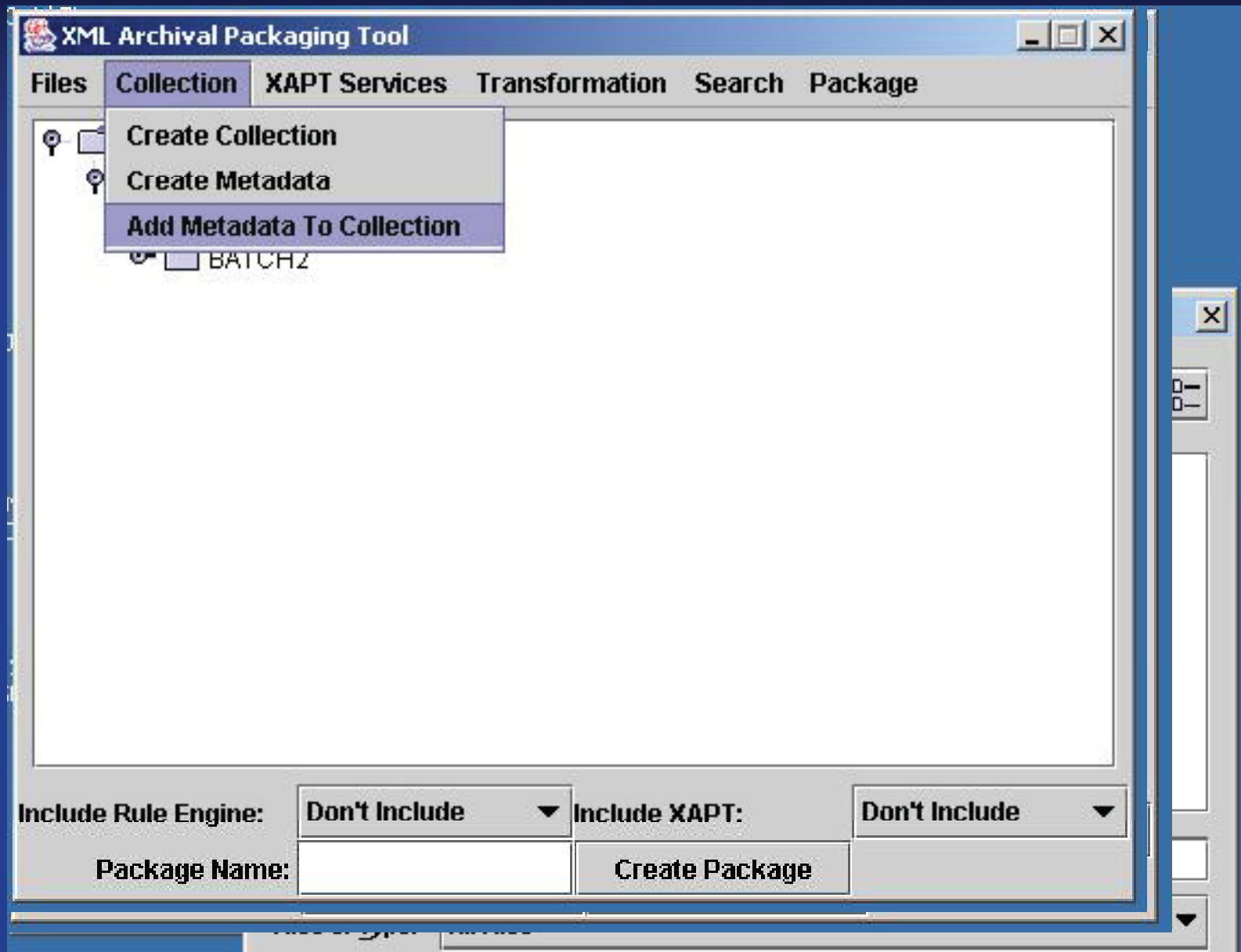
2. Import BATCH1 and BATCH2 into workspace



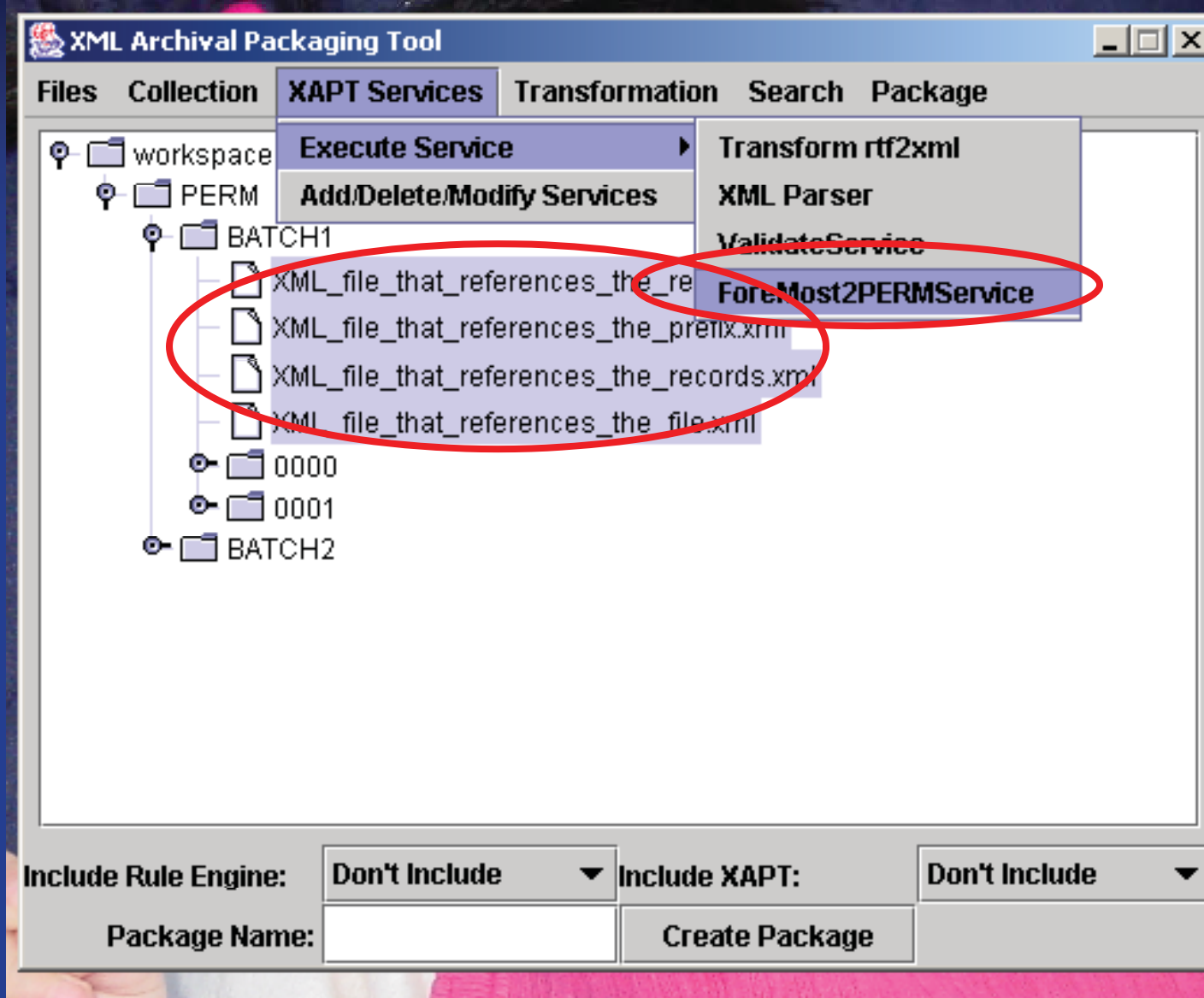
BATCH1 and BATCH2 metadata and contents inside XAPT workspace



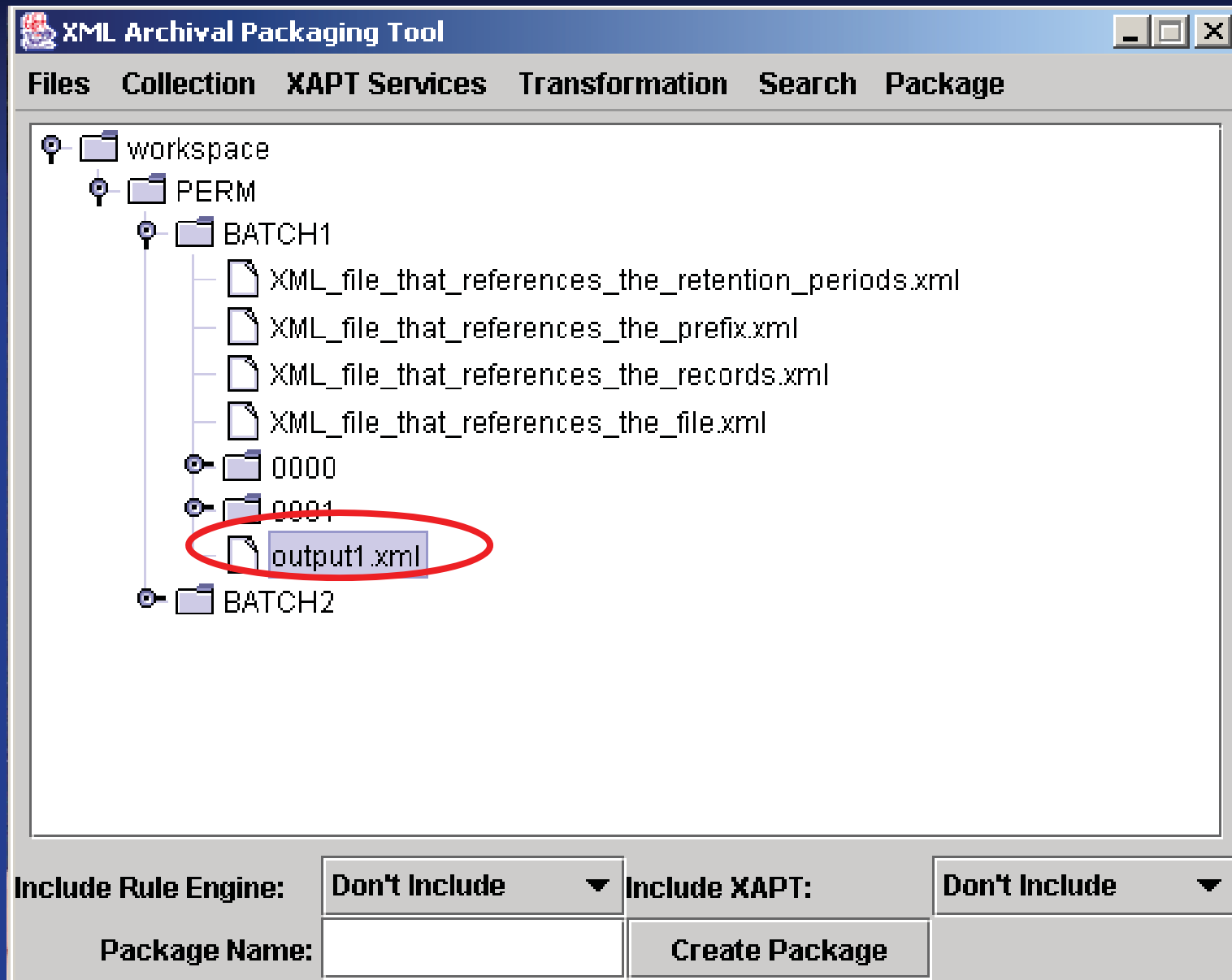
3. Create Collection Metadata



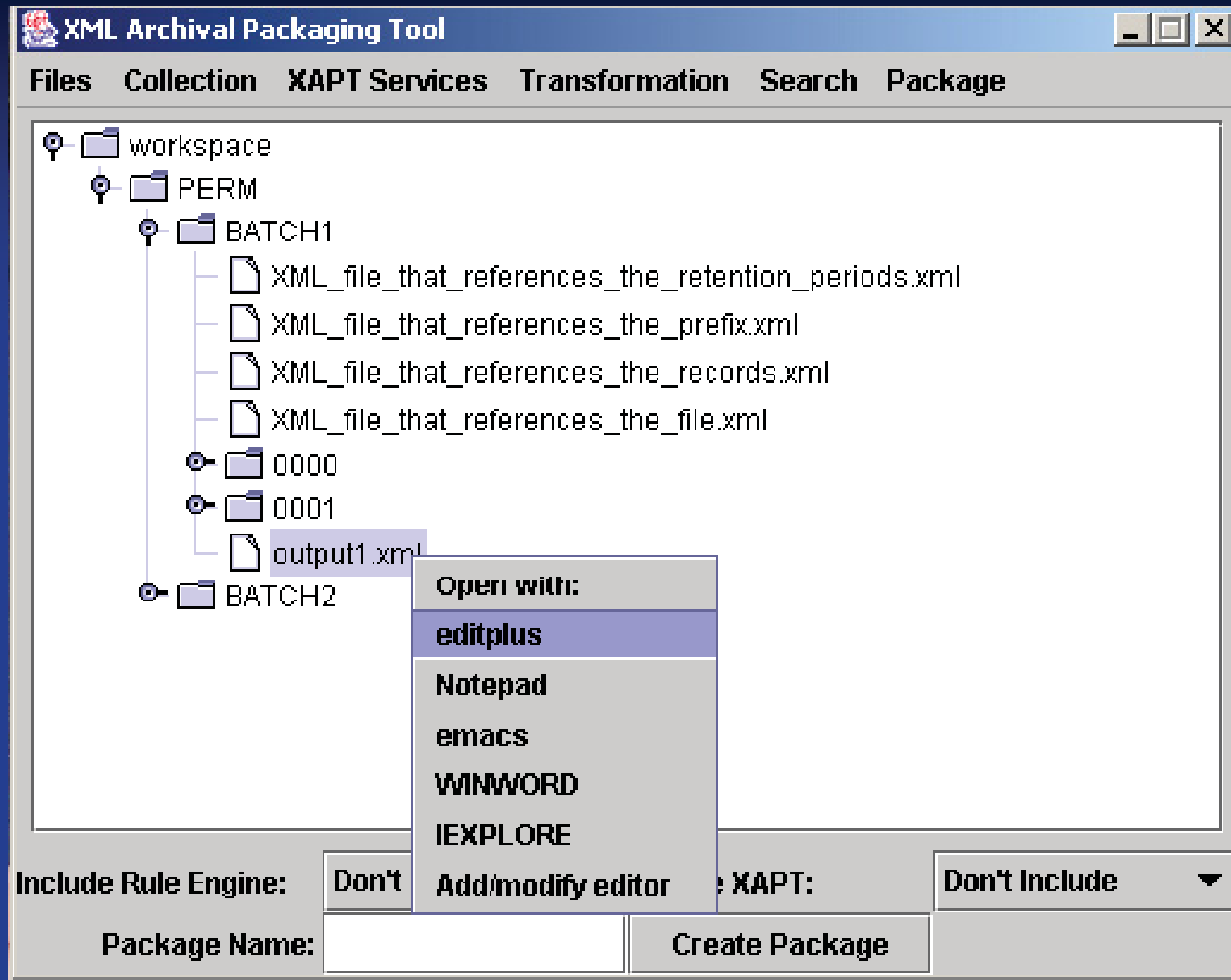
4. Consolidate BATCH1's Metadata Files into a PERM Format



PERM metadata shows up in workspace



Open PERM metadata file (DoDSTD1.xml)



C2.T2 = Record Folder Components

C2.T2.1.3 (Record Location)

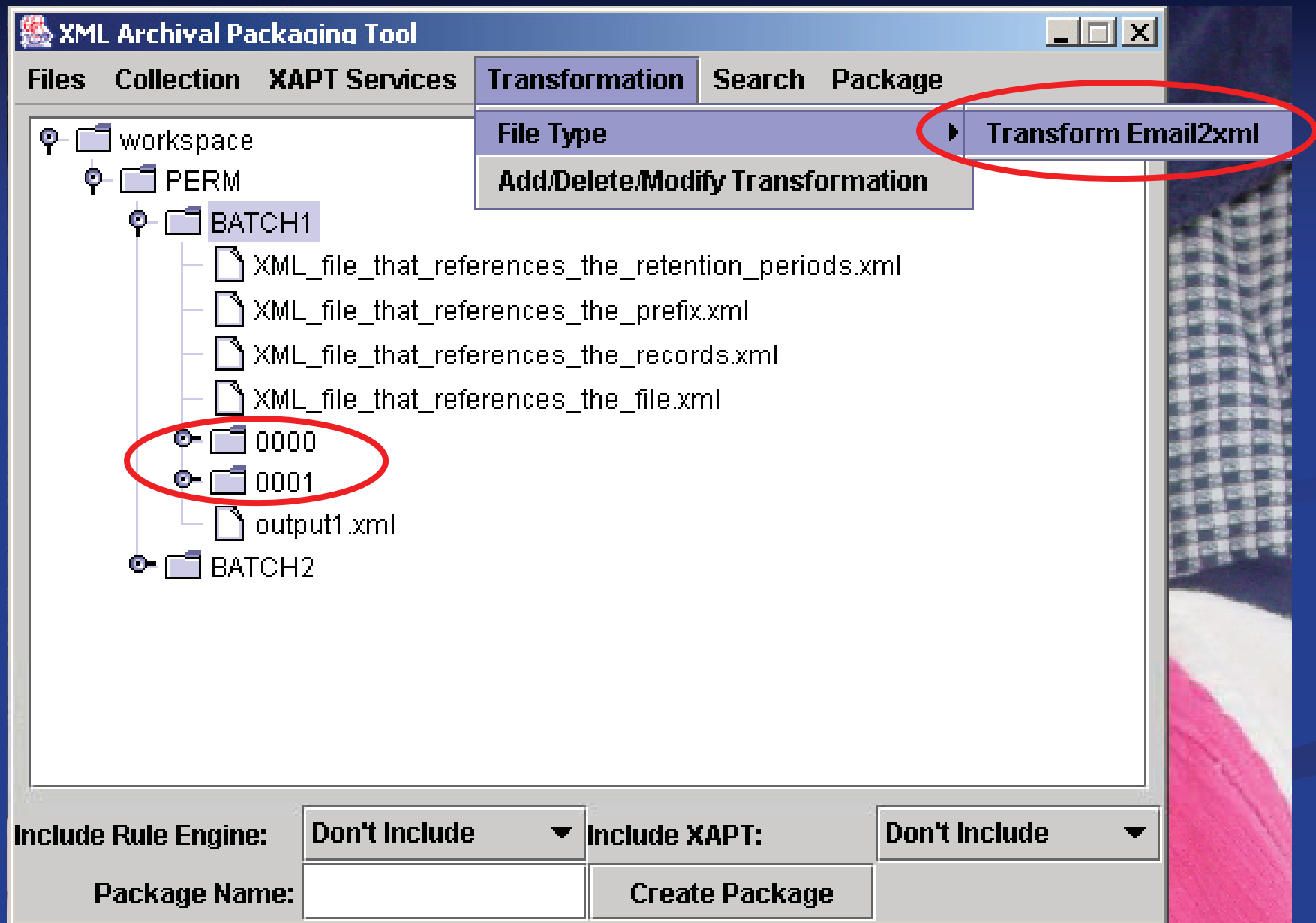
→ Linked to the data file 0001\70\00017036.doc

The screenshot shows an XML editor window titled "EditPlus - [D:\XAPT\Service\DoDSTD1.xml]". The main text area displays an XML document with the following structure:

```
69683 <C2.T1>
69684 <C2.T1.1>(GS5.02b) Administrative Subject File</C2.T1.1>
69685 <C2.T1.2>(GS5.02b) Administrative Subject File</C2.T1.2>
69686 <C2.T1.3>These records are used to support analysis, planning, procedure development, and activiti
69687 <C2.T1.4></C2.T1.4>
69688 <C2.T1.5>$AB</C2.T1.5>
69689 <C2.T1.6></C2.T1.6>
69690 <C2.T1.7>N</C2.T1.7>
69691 <C2.T1.8></C2.T1.8>
69692 <C2.T1.9></C2.T1.9>
69693 </C2.T1>
69694 <C2.T2>
69695 <C2.T2.1></C2.T2.1>
69696 <C2.T2.1.1>Administration (ACT+5)</C2.T2.1.1>
69697 <C2.T2.1.2>1000</C2.T2.1.2>
69698 <C2.T2.1.3>0001\70\00017036.doc</C2.T2.1.3>
69699 <C2.T2.1.4></C2.T2.1.4>
69700 <C2.T2.1.5></C2.T2.1.5>
69701 <C2.T2.1.6></C2.T2.1.6>
69702 <C2.T2.1.7>These records are used to support administrative analysis, planning, procedure developme
69703 </C2.T2>
69704 <C2.T3>
69705 <C2.T3.1>58323</C2.T3.1>
69706 <C2.T3.2></C2.T3.2>
69707 <C2.T3.3>Business Plan Activities.doc</C2.T3.3>
69708 <C2.T3.4>E</C2.T3.4>
69709 <C2.T3.5>.doc</C2.T3.5>
69710 <C2.T3.6>2002-11-01T00:00:00</C2.T3.6>
69711 <C2.T3.7>2002-11-01T00:00:00</C2.T3.7>
69712 <C2.T3.8></C2.T3.8>
69713 <C2.T3.9>Kinsella, Jim</C2.T3.9>
69714 <C2.T3.10>T0 FILE</C2.T3.10>
69715 <C2.T3.11>T0 FILE</C2.T3.11>
69716 <C2.T3.12>Department of Management and Budget, Management Services, Agency Services, Records Manage
69717 <C2.T3.13>0</C2.T3.13>
69718 <C2.T3.14></C2.T3.14>
69719 <C2.T3.15></C2.T3.15>
69720 <C2.T3.16>Manage records and information for state government in a cost effective and efficient ma
69721 </C2.T3>
69722 <C4.T1>
```

The record location path `<C2.T2.1.3>0001\70\00017036.doc</C2.T2.1.3>` is highlighted in blue. The left sidebar shows a file explorer view of the local drive, with the file `0001\70\00017036.doc` visible in the directory tree. The status bar at the bottom indicates the current position in the document: "In 69698 col 60 69920 00 PC REC INS READ".

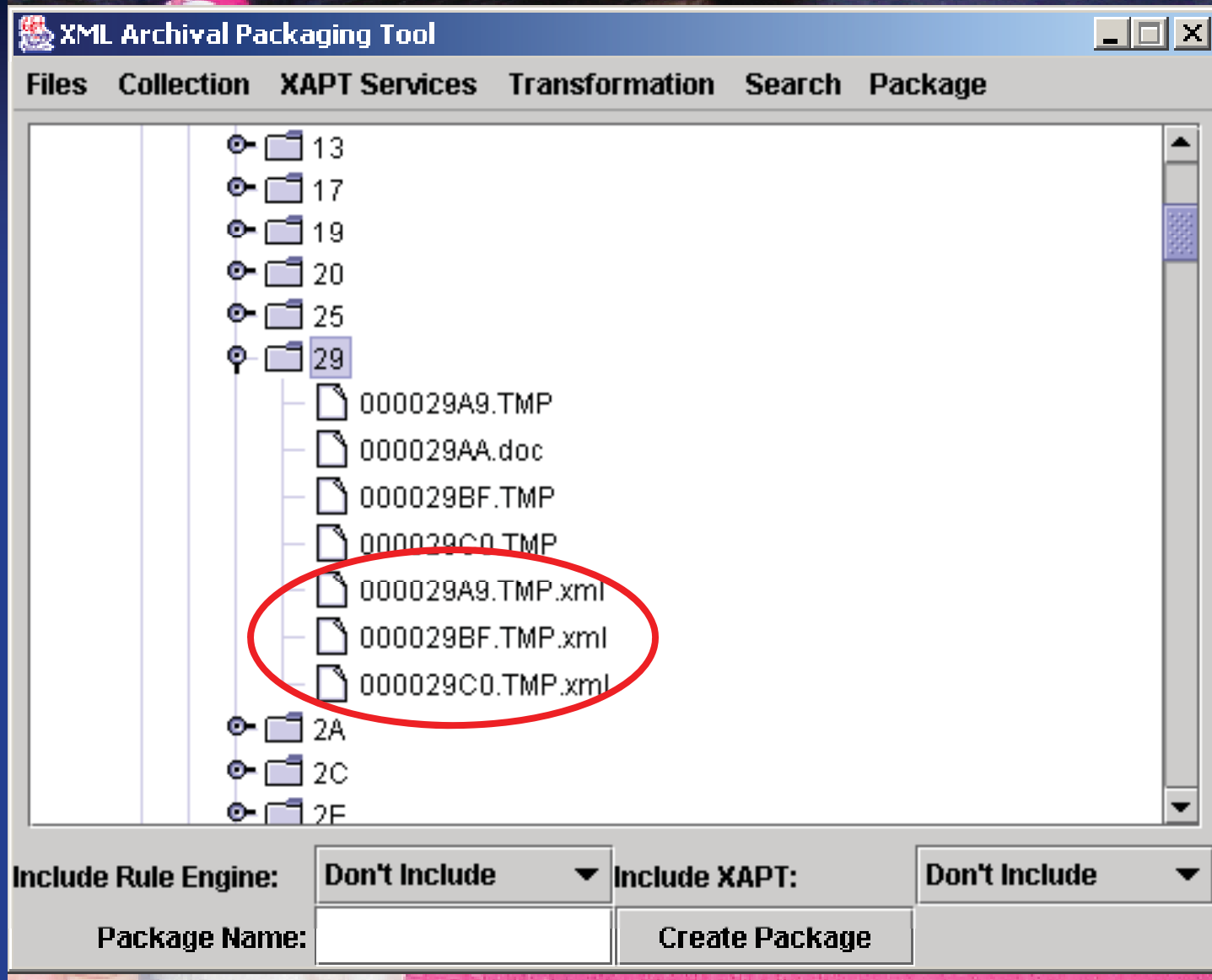
5. Bulk transformation of Email files (.tmp) in BATCH1 into .XML files



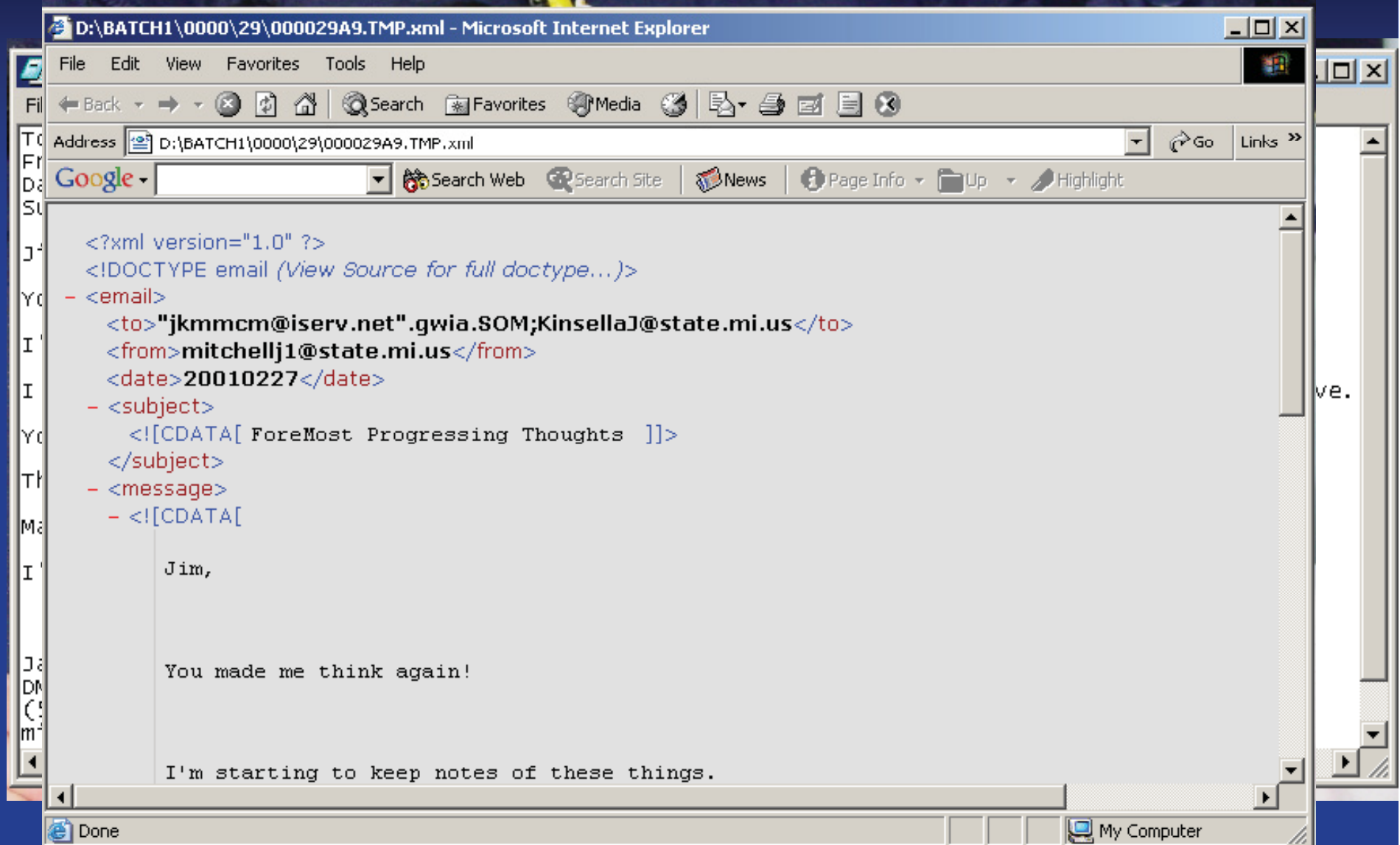
Conversion of all 602 files

```
C:\ XAPT - java XAPT
File no 579: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\E8\0000E820.TMP
File no 580: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\E8\0000E821.TMP
File no 581: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\E8\0000E824.TMP
File no 582: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\E8\0000E850.TMP
File no 583: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\E8\0000E86C.TMP
File no 584: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\E8\0000E89C.TMP
File no 585: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\E8\0000E8B2.TMP
File no 586: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\E8\0000E8B3.TMP
File no 587: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EAC7.TMP
File no 588: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EAC8.TMP
File no 589: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EAC9.TMP
File no 590: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EACA.TMP
File no 591: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EACD.TMP
File no 592: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EACF.TMP
File no 593: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EAD0.TMP
File no 594: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EAD1.TMP
File no 595: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EAD2.TMP
File no 596: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EA\0000EADE.TMP
File no 597: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EB\0000EB33.TMP
File no 598: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EB\0000EB34.TMP
File no 599: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EC\0000EC52.TMP
File no 600: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EC\0000EC53.TMP
File no 601: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EC\0000EC54.TMP
File no 602: C:\ArchivistsWorkbench\Michigan_DATA\BATCH1\0000\EF\0000EFA5.TMP
```

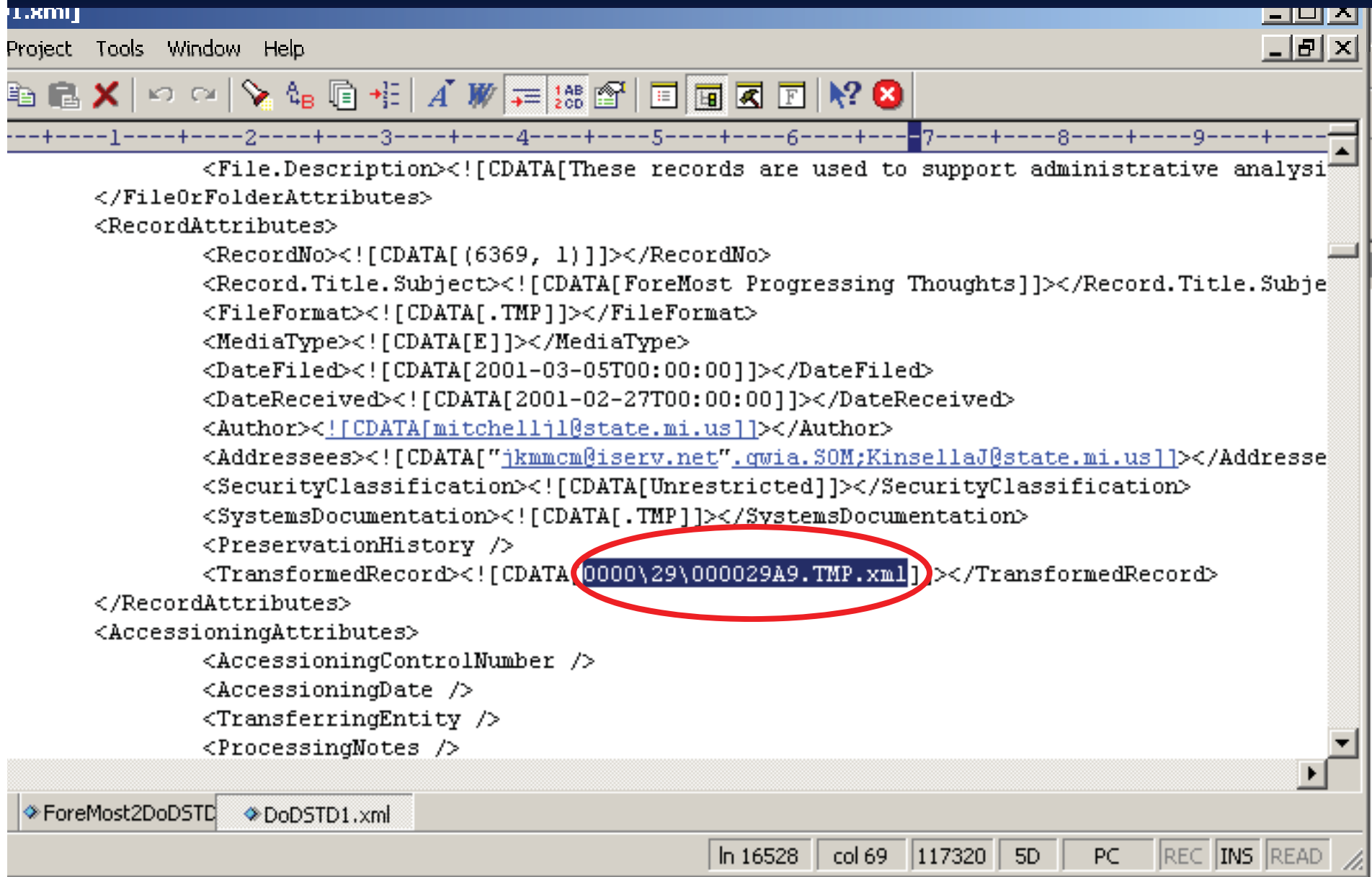
.TMP.xml files show up in the workspace



Viewing before and after: 000029A9.TMP and its transformed 000029A9.TMP.xml file



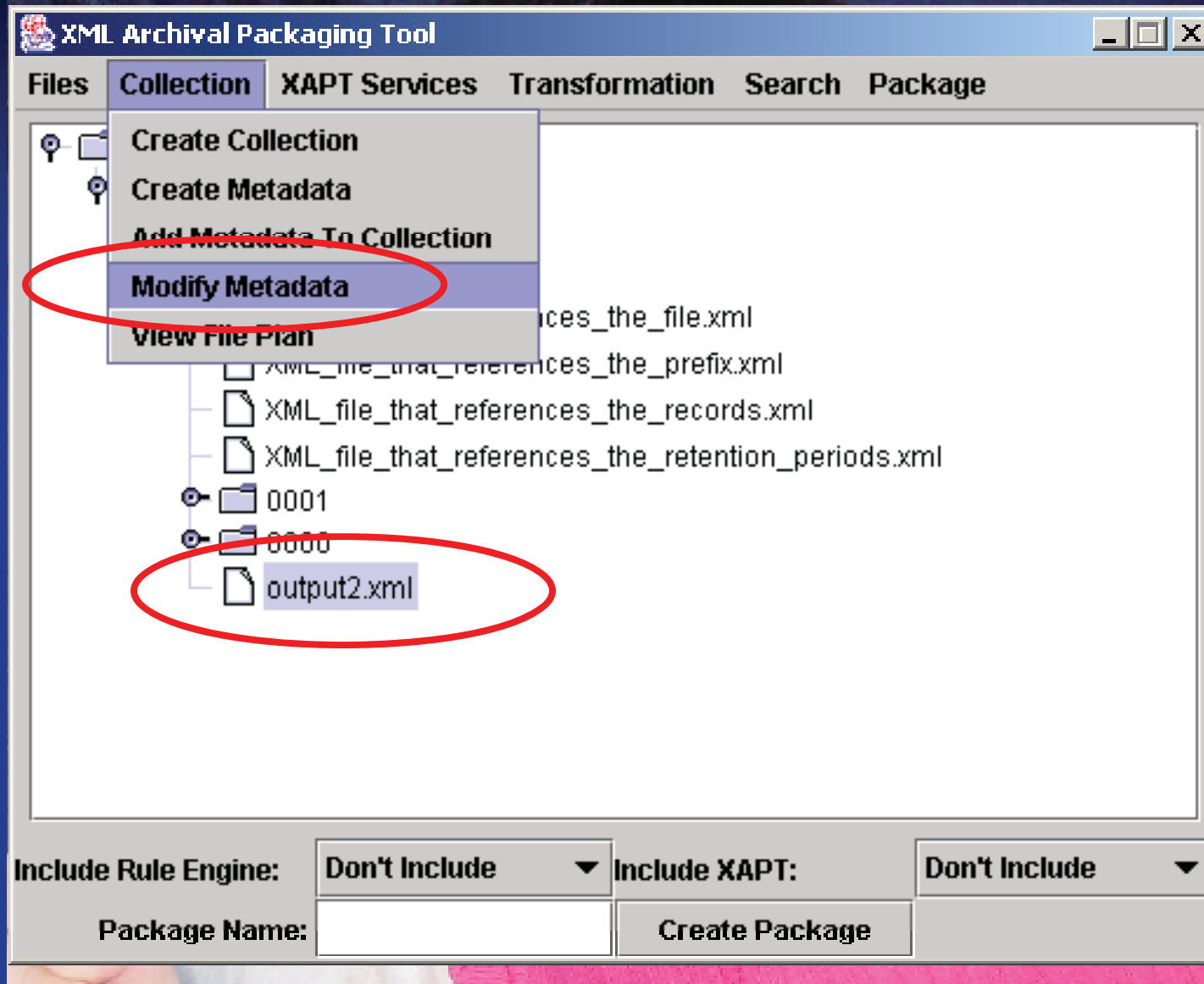
Linking to transformed record



```
<File.Description><![CDATA[These records are used to support administrative analysis  
</FileOrFolderAttributes>  
<RecordAttributes>  
  <RecordNo><![CDATA[(6369, 1)]]></RecordNo>  
  <Record.Title.Subject><![CDATA[ForeMost Progressing Thoughts]]></Record.Title.Subje  
  <FileFormat><![CDATA[.TMP]]></FileFormat>  
  <MediaType><![CDATA[E]]></MediaType>  
  <DateFiled><![CDATA[2001-03-05T00:00:00]]></DateFiled>  
  <DateReceived><![CDATA[2001-02-27T00:00:00]]></DateReceived>  
  <Author><![CDATA[mitchelljl@state.mi.us]]></Author>  
  <Addressees><![CDATA["jkmmcm@iserv.net".qwia.SOM;KinsellaJ@state.mi.us]]></Addresse  
  <SecurityClassification><![CDATA[Unrestricted]]></SecurityClassification>  
  <SystemsDocumentation><![CDATA[.TMP]]></SystemsDocumentation>  
  <PreservationHistory />  
  <TransformedRecord><![CDATA[0000\29\000029A9.TMP.xml]]></TransformedRecord>  
</RecordAttributes>  
<AccessioningAttributes>  
  <AccessioningControlNumber />  
  <AccessioningDate />  
  <TransferringEntity />  
  <ProcessingNotes />
```

The screenshot shows an XML editor window titled "1.xml" with a menu bar (Project, Tools, Window, Help) and a toolbar. The XML content is displayed in a text area with a ruler at the top. The value "0000\29\000029A9.TMP.xml" within the <TransformedRecord> tag is circled in red. The status bar at the bottom shows "ln 16528 col 69 117320 5D PC REC INS READ".

6. Modify Preservation Metadata

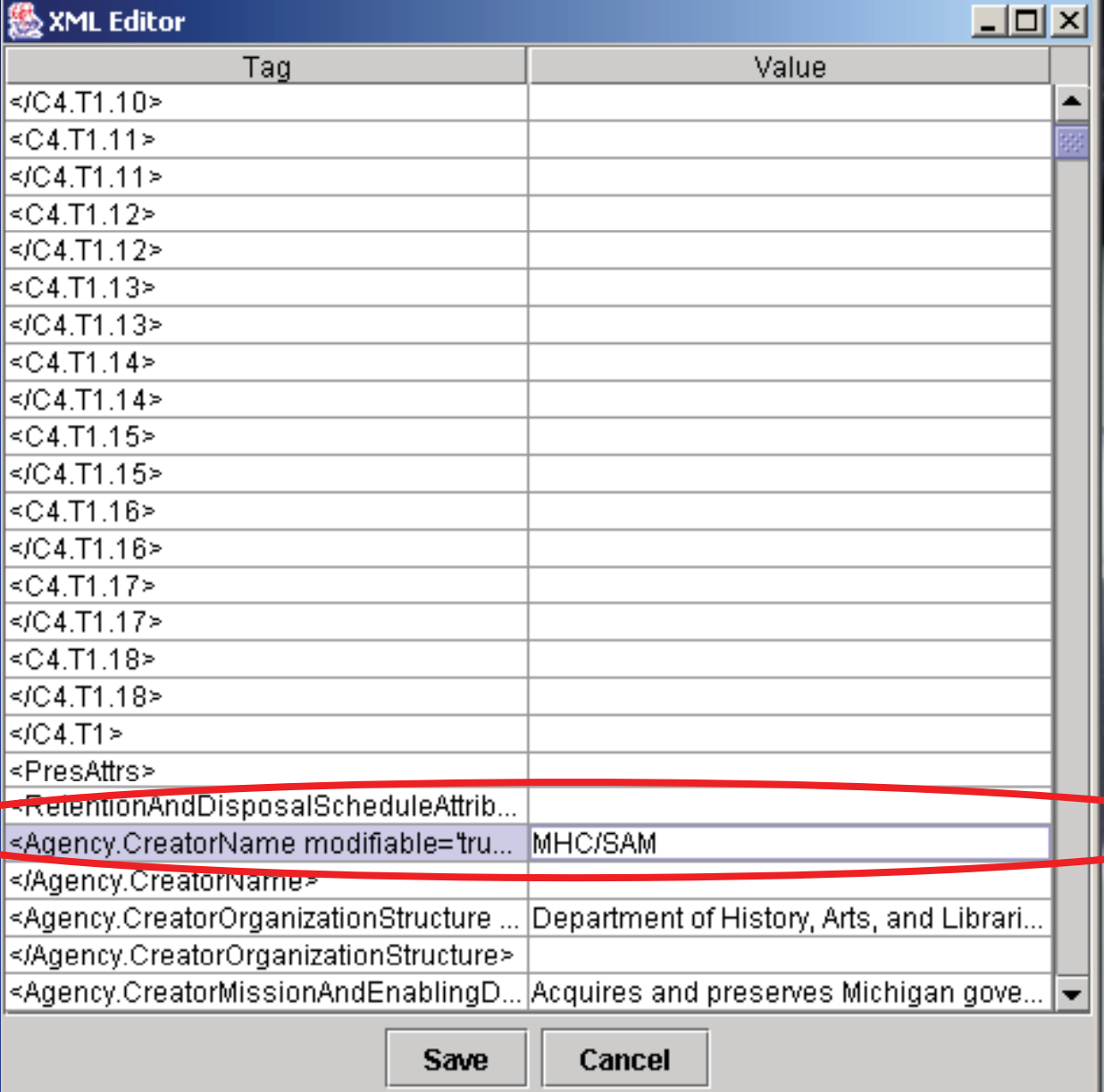


PERM Preservation Attributes

Preservation Attributes Used in the Preservation System			
Proposed Preservation Attributes for Archival Records	DoD 5015.2-STD (06/19/2002)	Michigan's ForeMost Field	Modifiable Preservation Attribute (Attribute in the Preservation System)
Retention and Disposal Schedule Attributes			
Agency/Creator Name	C2.T3.1.2 - "Originating Organization" - Mandatory	Prefix Name	Yes
Agency/Creator Organization Structure	N/A	Prefix Description (Primary)	Yes
Agency/Creator Mission and Enabling Document Citation	N/A	Prefix Description (Secondary)	
Record Series/Category Attributes			
Record Series #	C2.T1.2 - "Record Category Identifier" - Mandatory	File Description (Primary)	
Title/Subject	C2.T1.1 - "Record Category Name" - Mandatory	File Description (Primary)	
Description	C2.T1.3 - "Record Category Description" - Mandatory	File Description (Secondary)	
Retention Period	C2.T1.4 - "Disposition Instructions" - Mandatory	File Retention Period Id Name)	
Appraisal Value	C2.T1.6 - "Permanent Record Indicator" - Mandatory	File Retention Period Id	
Confidentiality Status	C2.T3.2 - "Supplemental Marking List" - Mandatory	File Security Level	
File or Folder Attributes			
File #	C2.T2.1.2 - "Folder Unique Identifier" - Mandatory	File	
Title/Subject	C2.T2.1.1 - "Folder Name" - Mandatory	File Subject (Primary)	
Description	N/A	File Subject (Secondary)	
Record Attributes			
Record #	C2.T3.1 - "Unique Record Identifier" - Mandatory	Document Number	No
Title/Subject	C2.T3.3 - "Subject or Title" - Mandatory	Document Subject	Yes
File Format	C2.T3.5 - "Format" -	Series (provides the file	Yes

	Mandatory	extension)	
Media Type	C2.T3.4 - "Media Type" - Mandatory	Document Type (Default Electronic or Default Non-Electronic)	Yes
Date Filed	C2.T3.6 - "Date Filed" - Mandatory	Document Date Filed	No
Date Received	C2.T3.8 - "Date Received" - Mandatory	Document Received Date	No
Author	C2.T3.9 - "Author or Originator" - Mandatory	From	No
Addressee(s)	C2.T3.10-11 - "Addressee(s) and Other Addressee(s)" - Mandatory	To	No
Security Classification	C2.T3.2 - "Supplemental Marking List" - Mandatory	Document Security	Yes
Systems Documentation (Hardware, OS, Software Name and Version)	N/A	Series (contains extension for file format)	Yes
Preservation History	N/A	N/A	Yes (append only)
Accessioning Attributes (Non-Mandatory Fields Populated by the Archival Repository)			
Accessioning Control Number	N/A	N/A	Yes
Accessioning Date	N/A	N/A	No
Transferring Entity	N/A	N/A	Yes
Processing Notes	N/A	N/A	Yes

... blue background indicates modifiable value



Tag	Value
</C4.T1.10>	
<C4.T1.11>	
</C4.T1.11>	
<C4.T1.12>	
</C4.T1.12>	
<C4.T1.13>	
</C4.T1.13>	
<C4.T1.14>	
</C4.T1.14>	
<C4.T1.15>	
</C4.T1.15>	
<C4.T1.16>	
</C4.T1.16>	
<C4.T1.17>	
</C4.T1.17>	
<C4.T1.18>	
</C4.T1.18>	
</C4.T1>	
<PresAttrs>	
<RetentionAndDisposalScheduleAttrib...	
<Agency.CreatorName modifiable='tru...	MHC/SAM
</Agency.CreatorName>	
<Agency.CreatorOrganizationStructure ...	Department of History, Arts, and Librari...
</Agency.CreatorOrganizationStructure>	
<Agency.CreatorMissionAndEnablingD...	Acquires and preserves Michigan gove...

Save Cancel

7. Extract File Plan for BATCH2 (in .XML)

The screenshot displays two overlapping windows. The background window is Microsoft Internet Explorer, showing the file plan for 'D:\XAPT\fileplan.xml'. The foreground window is the XML Archival Packaging Tool, with a context menu open over the 'View File Plan' option.

XML Archival Packaging Tool Context Menu:

- Create Collection
- Create Metadata
- Add Metadata To Collection
- Modify Metadata
- View File Plan

XML File Plan Content:

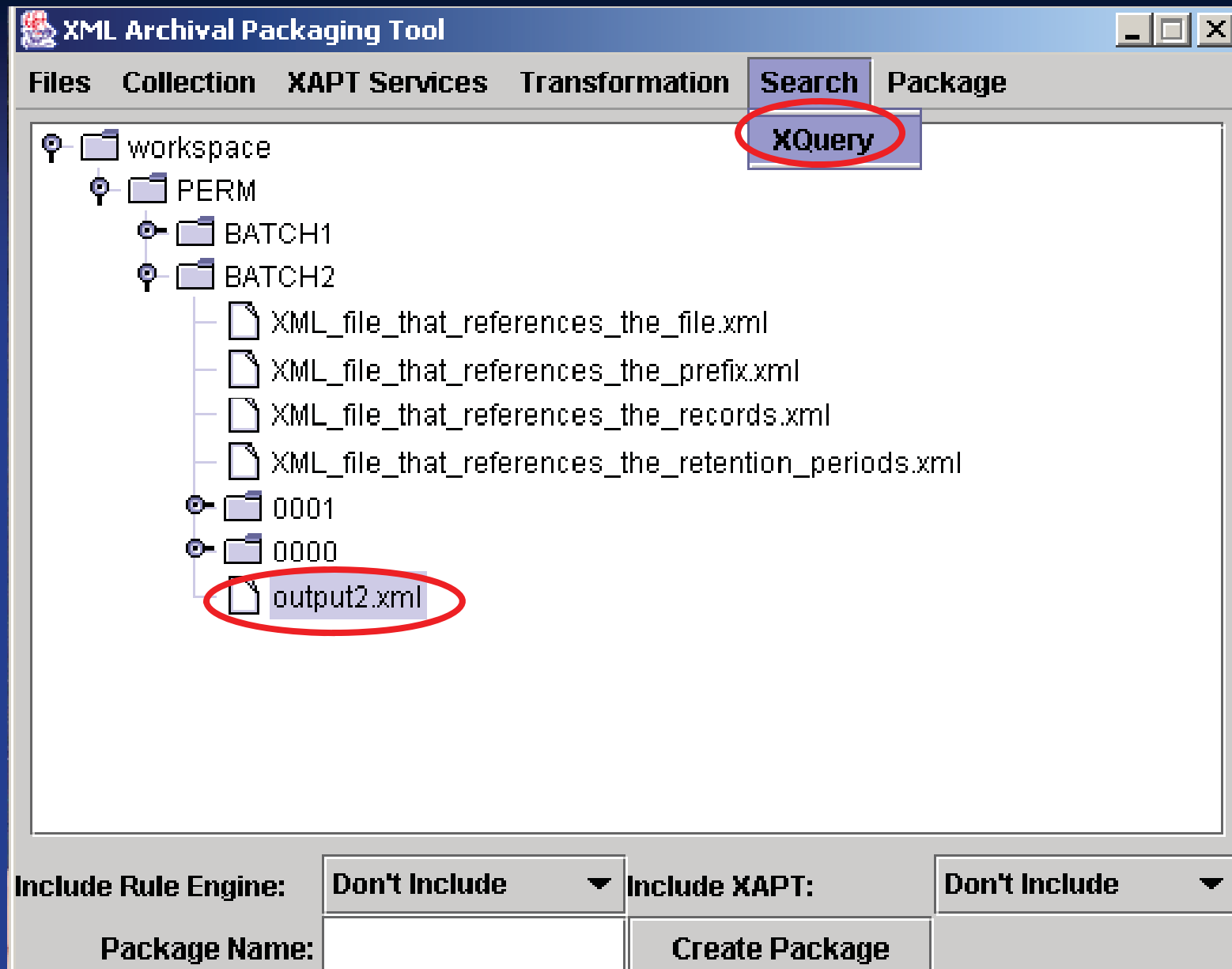
```
<?xml version="1.0" ?>
- <prefix name="MHC/SAM">
- <primary_file name="6600 [Records Management (Act+5)]">
- <secondary_file name="6600-10 [Disposals and Transmittals (Act+5)]">
  <document>0000\FF\0000FF2E.098</document>
  <document>0000\FF\0000FF2F.KEY</document>
  <document>0000\FF\0000FF30.KEY</document>
  <document>0000\FF\0000FF33.doc</document>
  <document>0000\FF\0000FF49.doc</document>
  <document>0001\3D\00013DED.dot</document>
  <document>0001\5B\00015B16.doc</document>
  <document>0001\5B\00015BEB.txt</document>
  <document>UUU1\5B\UUU15BEG.txt</document>
  <document>0001\5B\00015BED.822</document>
  <document>0001\5C\00015C3E.txt</document>
  <document>0001\5C\00015C41.txt</document>
  <document>0001\60\00016010.txt</document>
  <document>0001\60\00016011.txt</document>
  <document>0001\60\00016012.txt</document>
  <document>0001\60\00016013.txt</document>
  <document>0001\60\00016014.txt</document>
  <document>0001\63\000163CA.txt</document>
  <document>0001\63\000163CB.PDF</document>
  <document>0001\63\000163CC.822</document>
  <document>0001\64\0001643B.txt</document>
  <document>0001\64\0001643C.txt</document>
</secondary_file>
+ <secondary_file name="6600-20 [Retention and Disposal Schedules (Act+5)]">
+ <secondary_file name="6600-30 [Versatile (Act+5)]">
  <document>0001\53\000153F7.txt</document>
  <document>0001\53\000153F8.txt</document>
</secondary_file>
</primary_file>
</prefix>
```

XML Archival Packaging Tool Interface:

- Buttons: Include XAPT: (dropdown), Don't Include (dropdown), Create Package
- File list: references_the_file.xml, references_the_prefix.xml, references_the_records.xml, references_the_retention_periods.xml

Taskbar: Start button, My Computer, and several application icons including Comm..., XAPT, Tomcat, XML Ar..., D:\XAP..., Perm's..., Docum..., and D:\XA... The system clock shows 11:59 AM.

8. Querying the PERM Metadata



Find all records where the addressee contains 'Caryn' or 'Wojcik'
C2.T3 = Record Metadata Components (C2.T3.10 = "Adressee(s)")

The screenshot shows a Microsoft Internet Explorer window displaying an XML document. The main content is an XQuery result, which is a list of XML records. The first record is expanded, showing its metadata components: <C2.T1>, <C2.T2>, <C2.T3>, and <C4.T1>. The <C2.T3> component is further expanded to show <PresAttrs>, which contains the text: <code>//record[C2.T3/C2.T3.10[. &= 'Caryn Wojcik']]</code>. A red oval highlights this query string in the XQuery dialog box. The dialog box also shows the result of the query: <code>=====> 4 hit(s).</code>

```
<?xml version="1.0" ?>
- <xqe:result query="//record[ C2.T3/C2.T3.10[ . &= 'Caryn Wojcik' ] ]"
  xmlns:xqe="www.fatdog.com/XQEngine.html">
- <record>
+ <C2.T1>
+ <C2.T2>
+ <C2.T3>
+ <C4.T1>
+ <PresAttrs>
  //record[ C2.T3/C2.T3.10[ . &= 'Caryn Wojcik' ] ]
</record>
- <record>
+ <C2.T1>
+ <C2.T2>
+ <C2.T3>
+ <C4.T1>
+ <PresAttrs>
</record>
- <record>
+ <C2.T1>
+ <C2.T2>
+ <C2.T3>
+ <C4.T1>
+ <PresAttrs>
</record>
- <record>
+ <C2.T1>
+ <C2.T2>
+ <C2.T3>
+ <C4.T1>
+ <PresAttrs>
</record>
</xqe:result>
```

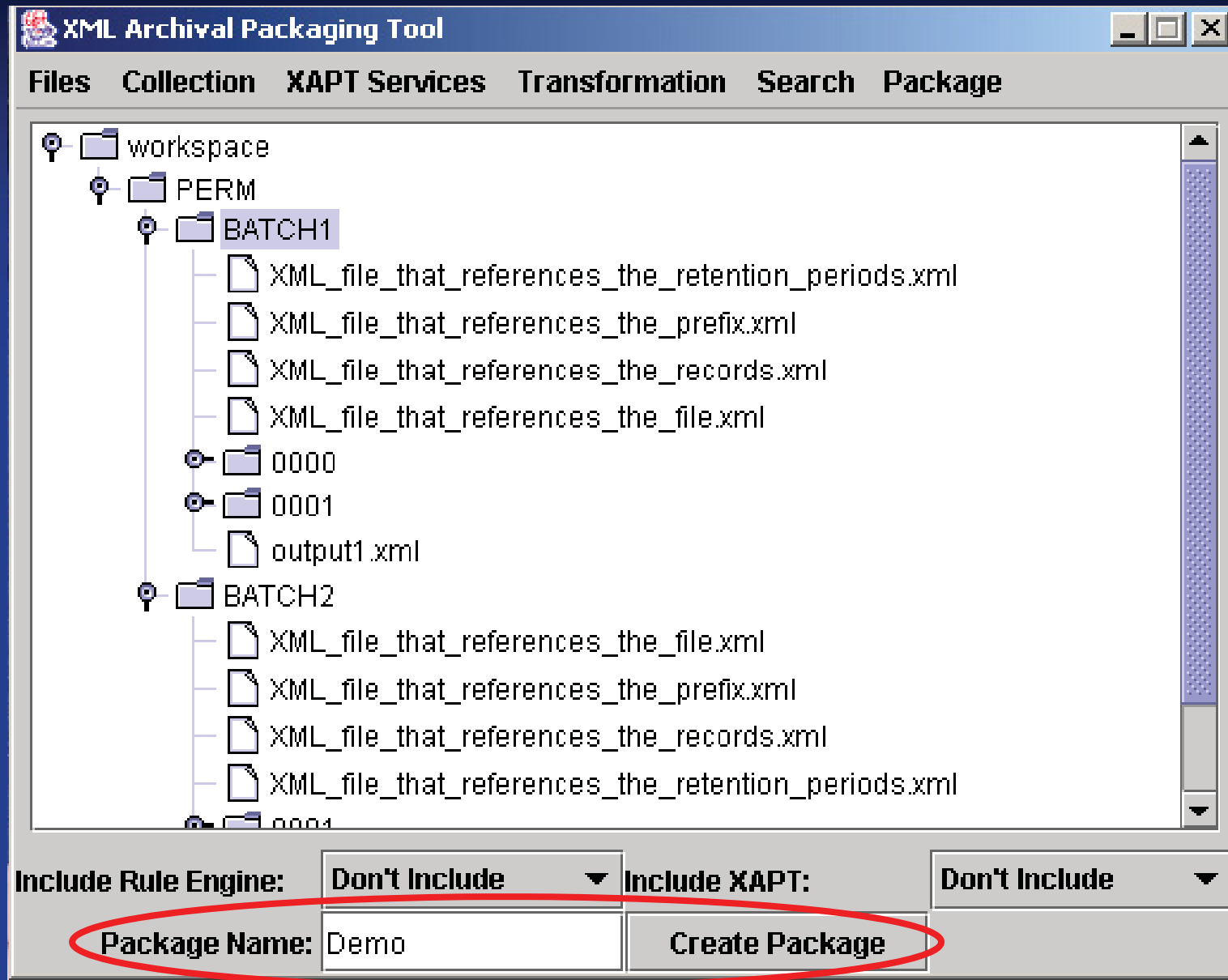
Retrieve the first one only

The screenshot shows a Microsoft Internet Explorer window displaying an XML document. The document is titled "C:\ArchivistsWorkbench\XAPT\XQueryResult.xml" and contains the following XML structure:

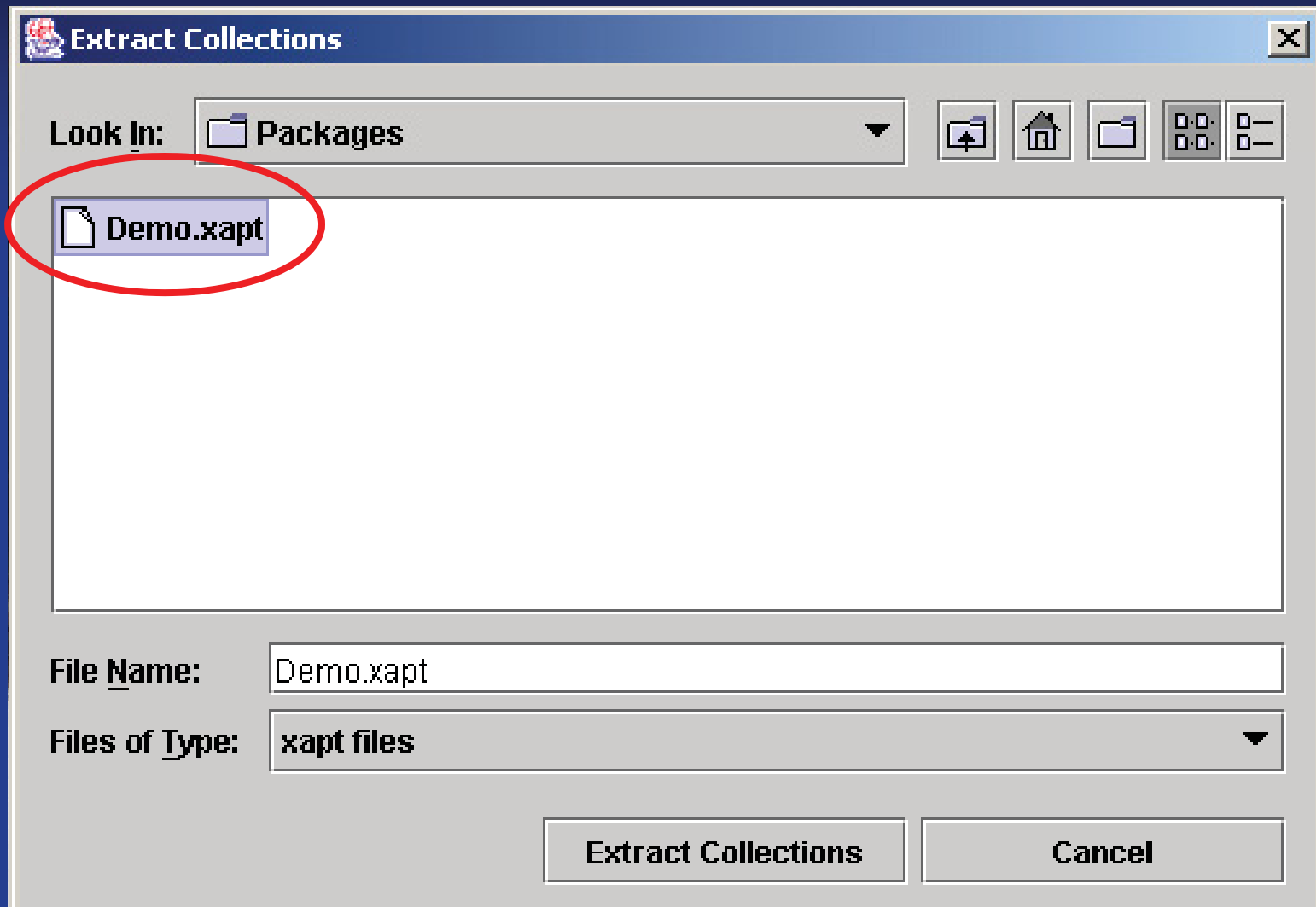
```
<?xml version="1.0" ?>
- <xqe:result query="//record[ C2.T3/C2.T3.10[ . &= 'Caryn Wojcik' ] ][1]"
  xmlns:xqe="www.fatdog.com/XQEngine.html">
- <record>
+ <C2.T1>
+ <C2.T2>
- <C2.T3>
  <C2.T3.1>(<?
  <C2.T3.2 />
  <C2.T3.3>El
  <C2.T3.4>E<
  <C2.T3.5>.T
  <C2.T3.6>2002-02-08T00:00:00</C2.T3.6>
  <C2.T3.7>2000-10-11T00:00:00</C2.T3.7>
  <C2.T3.8 />
  <C2.T3.9>Wojcik, Caryn</C2.T3.9>
  <C2.T3.10>Wojcik, Caryn</C2.T3.10>
  <C2.T3.11>Wojcik, Caryn</C2.T3.11>
  <C2.T3.12>Department of History, Arts, and Libraries,
    Michigan Historical Center, State Archives of
    Michigan.</C2.T3.12>
  <C2.T3.13>31002</C2.T3.13>
  <C2.T3.14 />
  <C2.T3.15 />
  <C2.T3.16>Acquires and preserves Michigan government
    records and manuscripts that document Michigan history
    for access by researchers.</C2.T3.16>
  </C2.T3>
+ <C4.T1>
+ <PresAttrs>
</record>
</xqe:result>
```

An XQuery dialog box is overlaid on the XML document. The dialog box has a title bar "XQuery" and a text input field labeled "Enter Query:". The query entered in the field is: `//record[C2.T3/C2.T3.10[. &= 'Caryn Wojcik']][1]`. Below the input field, the result of the query is displayed: `//record[C2.T3/C2.T3.10[. &= 'Caryn Wojcik']][1] =====> 1 hit(s).`. The query input field and the result text are both circled in red.

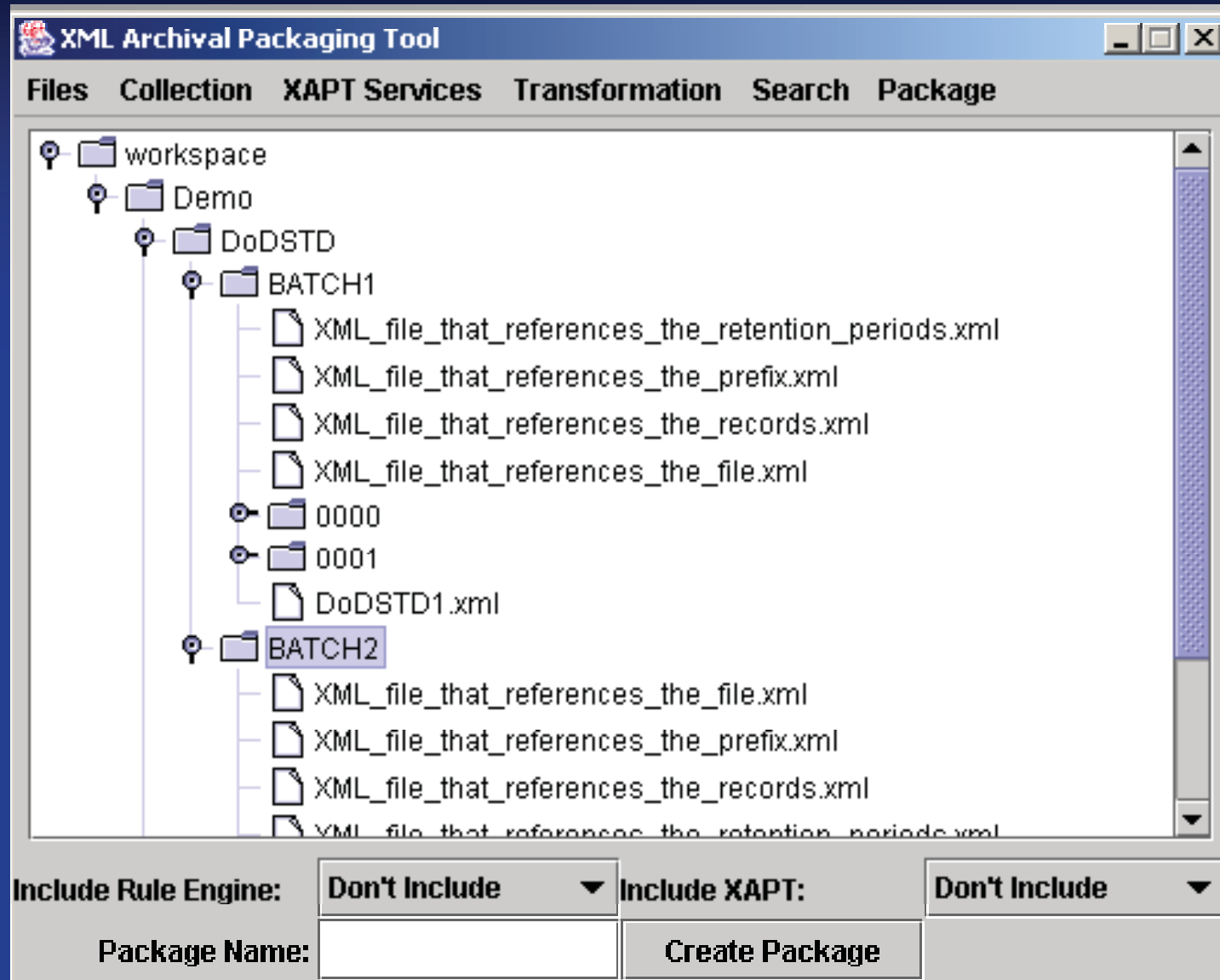
9. Create “Demo” package archive



10. Extract the collections from the Demo.xapt package:



→ BATCH1 and BATCH2 are reinstated into XAPT



Next Steps

ITERATIVE PROCESS:

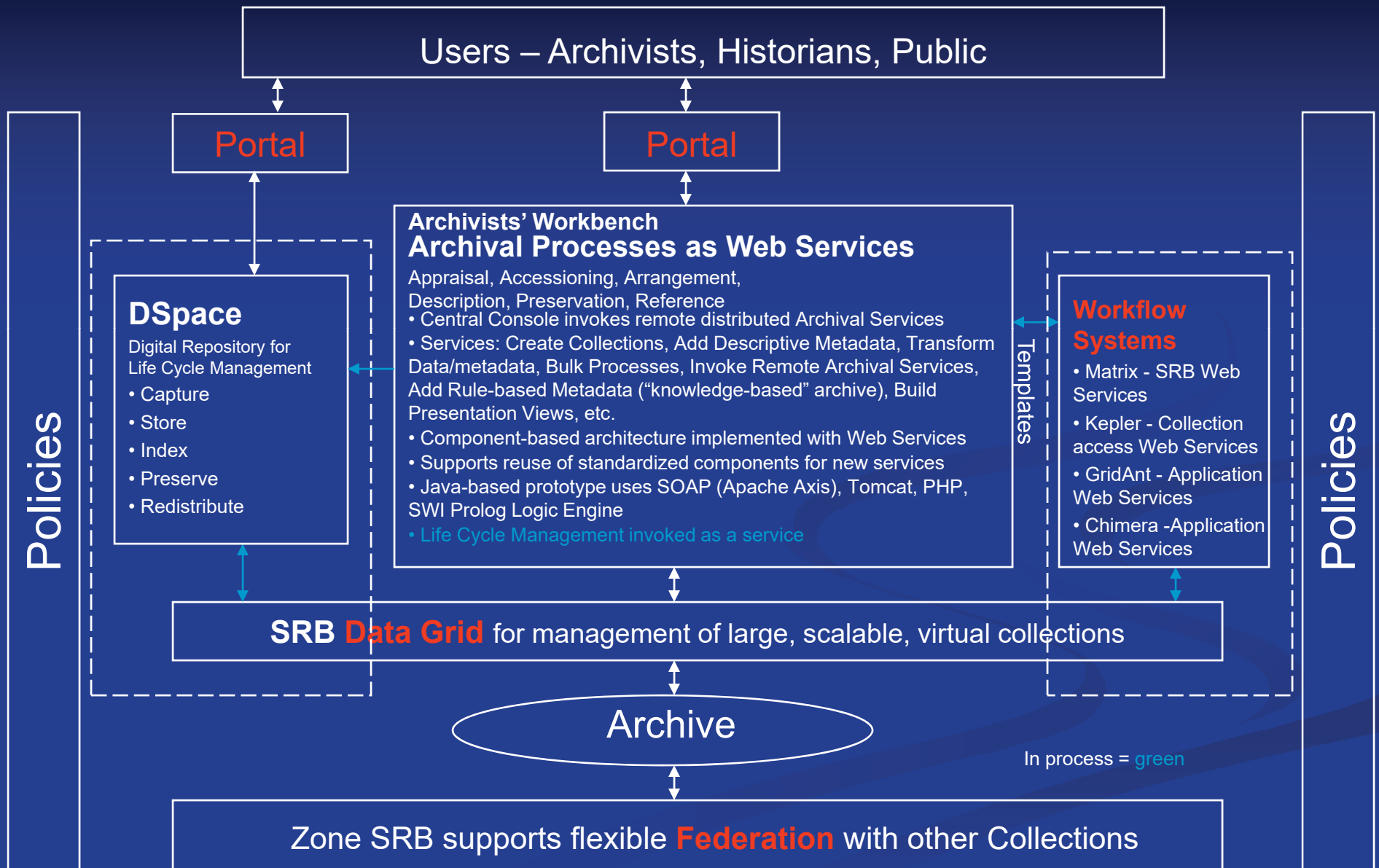
- Testing additional functional requirements
- Modifying functional requirements accordingly

- Proof of interoperability
 - Reloading the records and their associated preservation system attributes into the the original RMA repository
 - Loading the records and associated attributes into a different RMA

Additional Information

- Archivists' Workbench:
 - <http://www.sdsc.edu/NHPRC>
- PERM project:
 - <http://www.sdsc.edu/PERM>

SDSC Prototype Archivists' Workbench



Framework Components

Archivists' Workbench
Archival Processes as Web Services

- **Portal Technology**

- OGCE: NMI Middleware -- provide the Grid portal community with sharable portlet libraries that utilize Grid technologies.

Workflow Systems

Data Grids & Federation

Framework Components

Archivists' Workbench
Archival Processes as Web Services

Portal Technology

■ **Workflow Systems**

Data Grids & Federation

Senate Collection Example

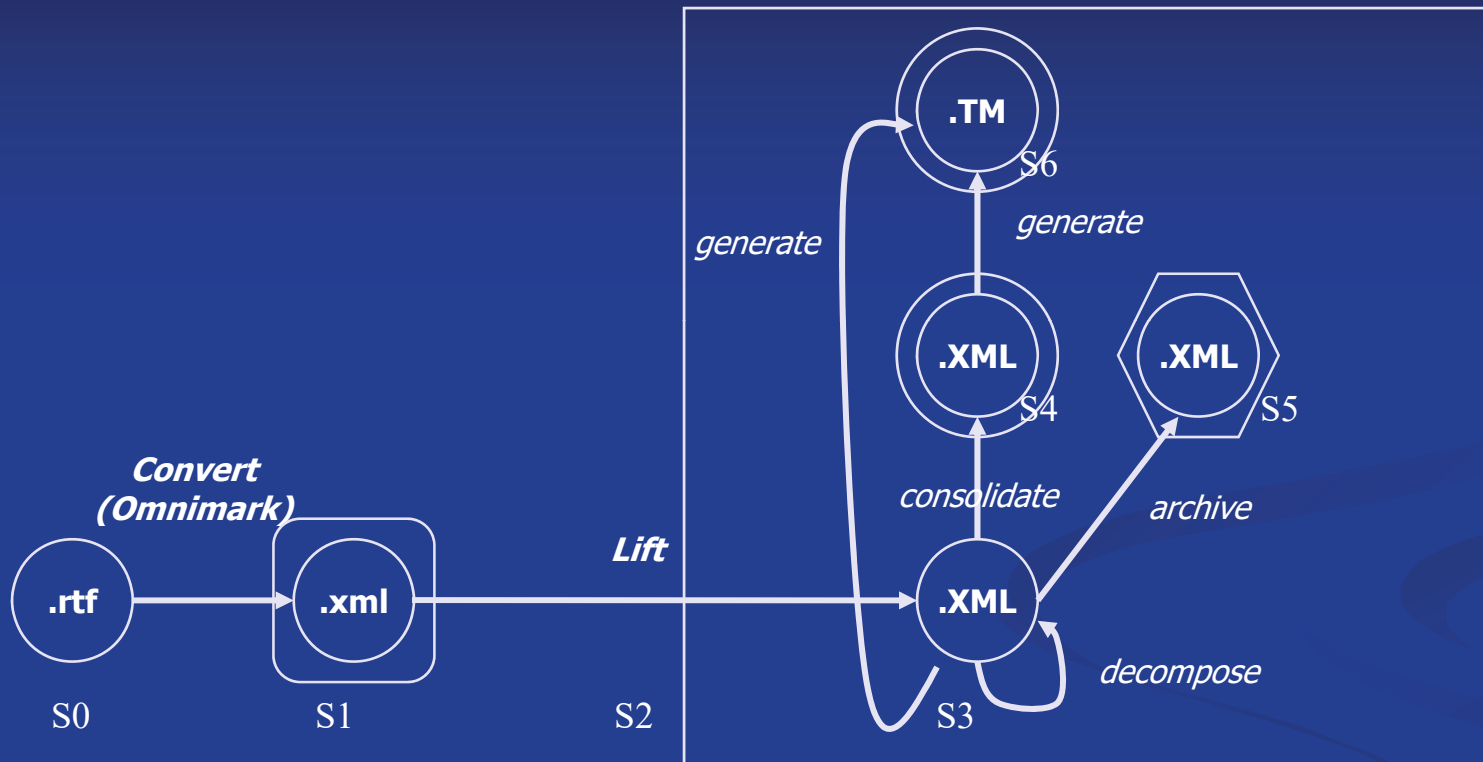
- ... the XML can be *lifted* from the *presentation* level:

```
<p bold="off">**** S. 345</p>
<p align="right" bold="off">DATE INTRODUCED: 02/03/1999</p>
<p bold="off">SPONSOR: Allard</p>
<p align="center" bold="off" italic="off">OFFICIAL TITLE</p>
<p bold="off" italic="off">A bill to amend the Animal Welfare Act to remove the lim\
itation that permits interstate movement of live birds, for the purpose of fighting\
, to States in which animal fighting is lawful.</p>
<p align="center" bold="off" italic="off">LATEST STATUS</p>
<p><string>Feb 3, 1999&tab;Read twice and referred to the Committee on Agriculture\
.</string></p>
<p></p>
```

- ... to the *information* level:

```
<bill name="S.345">
  <committees>
    <committee>SENATE: AGRICULTURE</committee>
  </committees>
  <date_introduced>02/03/1999</date_introduced>
  <latest_status_list>
    <latest_status> <ls_date>Feb 3, 1999</ls_date>
                    <ls_txt>Read twice and referred to the Committee on Agriculture</ls_txt>
  </latest_status>
</latest_status_list>
  <official_title>A bill to amend the Animal Welfare Act to remove the limitation that permits interstate movement of live birds, for
the purpose of fighting, to States in which animal fighting is lawful.</official_title>
  <sponsor>Allard, Wayne [CO]</sponsor>
</bill>
```

Ingestion Network: Y2K Example



Legend (stages):

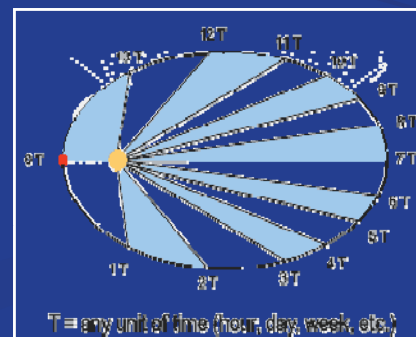
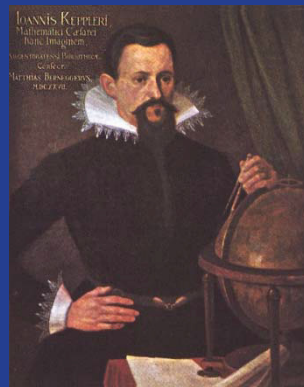




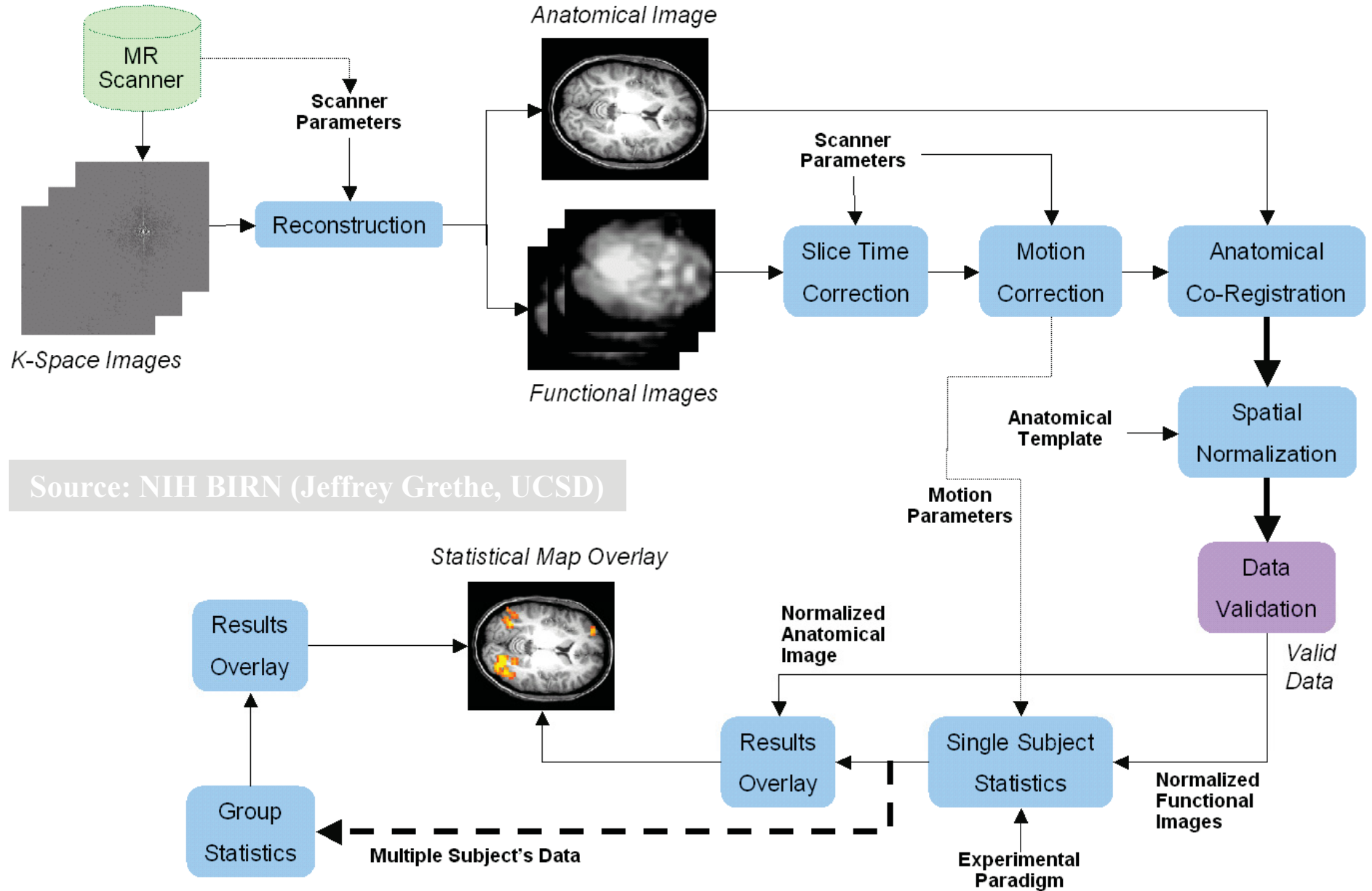
Workflow Systems

- Matrix - SRB Web Services
- Kepler - Collection access Web Services
- GridAnt - Application Web Services
- Chimera - Application Web Services

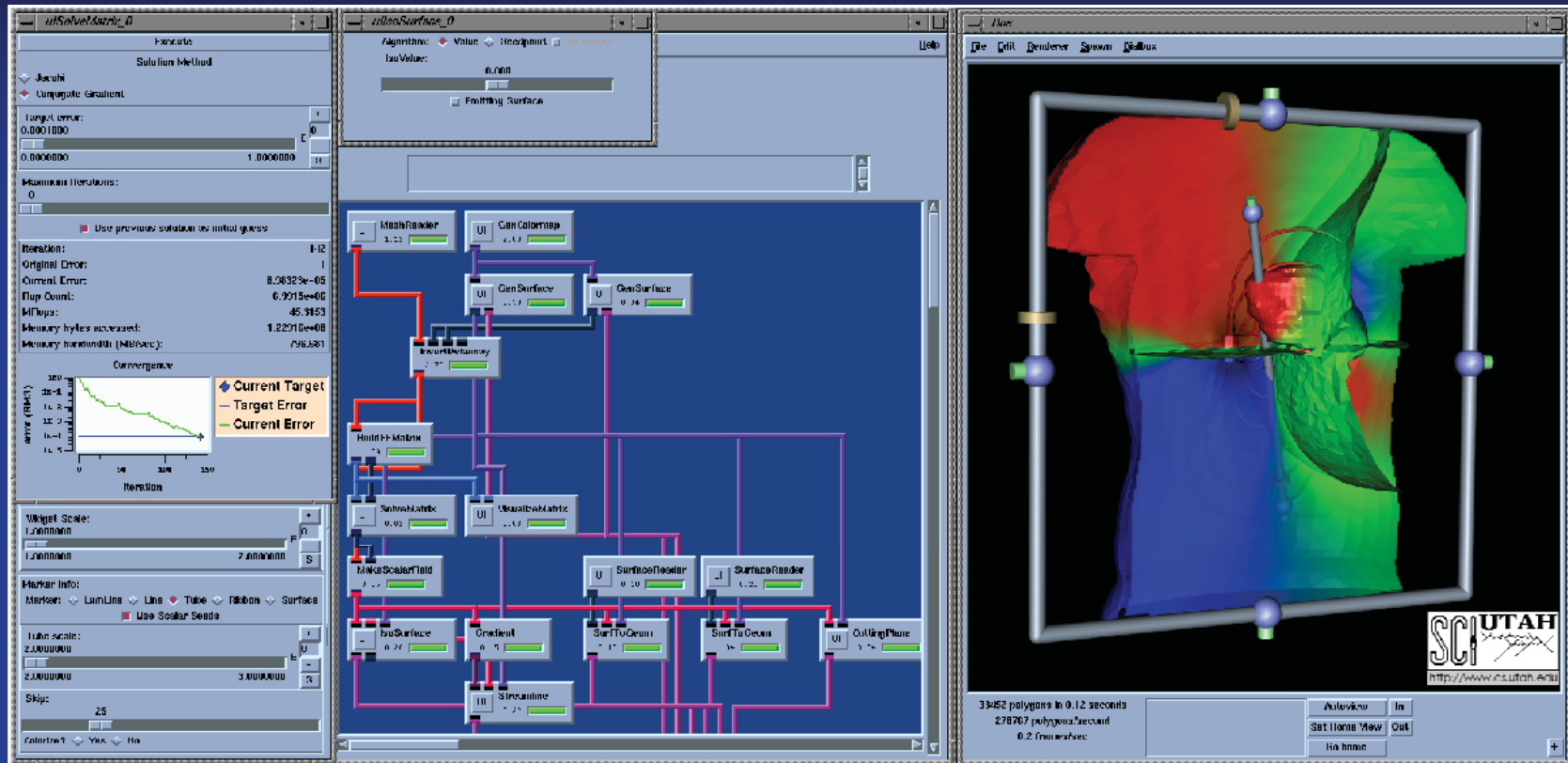
Kepler: Grid-Enabled Workflows



Functional MRI Analysis Workflow

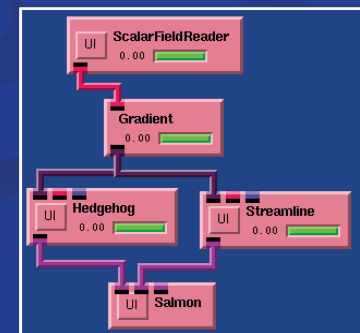


SCIRun: Problem Solving Environments for Large-Scale Scientific



- SCIRun: PSE for interactive construction, debugging, and steering of large-scale scientific computations
- New collaboration under Kepler/SDM
- Component model, based on generalized data flow programming

Steve Parker (cs.utah.edu)



The KEPLER GUI: Vergil (Steve Neuendorffer, Ptolemy II)

The screenshot shows the KEPLER GUI interface. On the left is a library pane with folders: utilities, director library, actor library, more libraries, and user library. A red arrow points from a text box at the bottom to the 'utilities' folder. The main workspace contains a workflow diagram with the following components and descriptions:

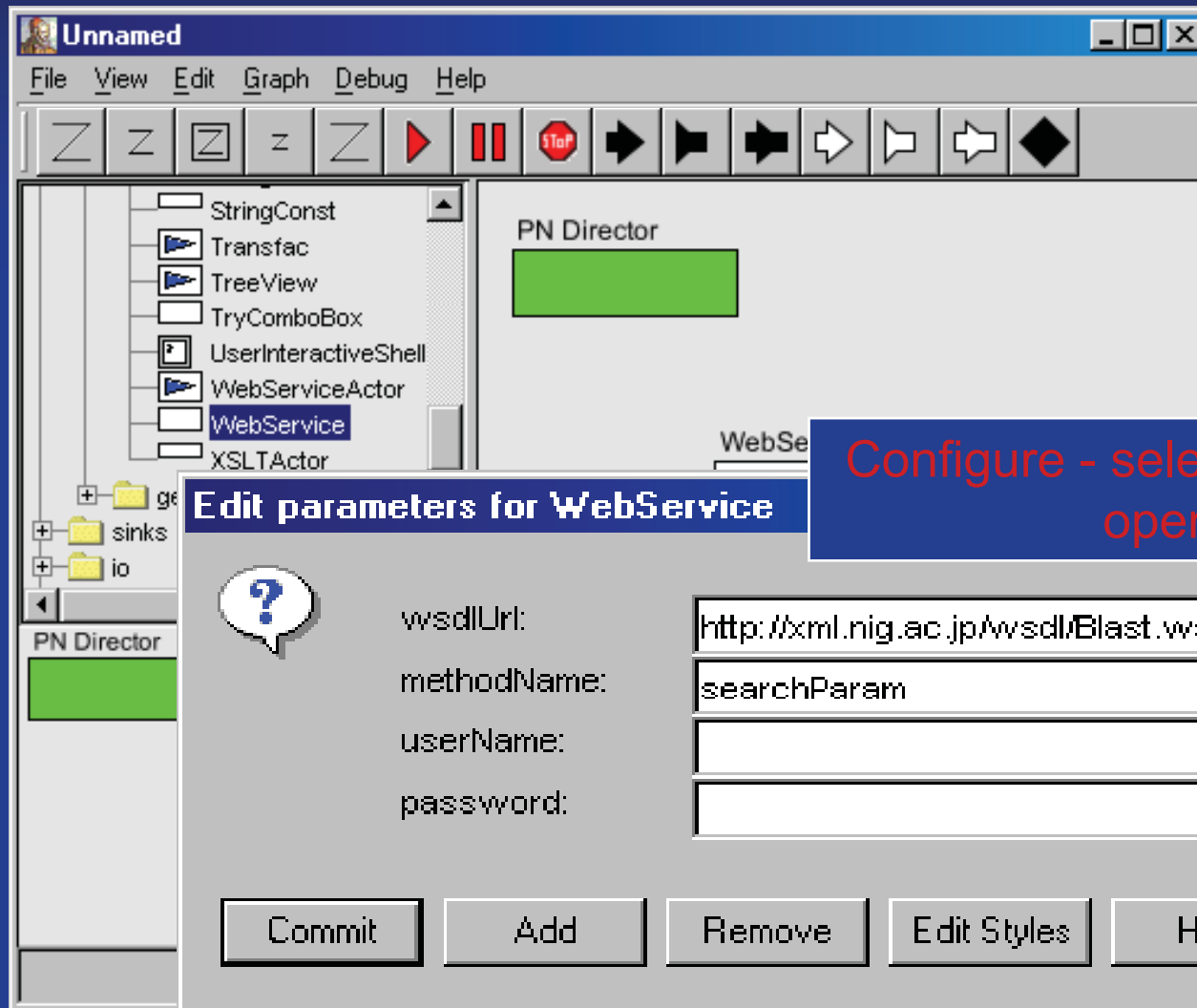
- PN Director**: Promoter identification workflow (PIW) aims at constructing models of transcription factor binding sites to identify co-regulated genes, starting from microarray data.
- FileReader for AccessNumbers**: Double click the File Reader to change specify your access numbers to be investigated. The access number should occupy one line each. Each access number will be investigated.
- Gene Sequence Processing**: This loop executes "Gene Sequence Processing" for each gene entered in "GeneAccessNumber List".
- FileReader**: Double-click to change the ClustalW filename.
- ClustalW_Remote**: Shows the physical alignment of multiple gene sequences. Uses DDBL-ClustalW Multiple Alignment Tool.
- ClustalW Results Display**: Displays the results of the ClustalW alignment.

The workflow diagram shows the following flow: FileReader for AccessNumbers feeds into Gene Sequence Processing and FileReader. Gene Sequence Processing feeds into ClustalW_Remote. ClustalW_Remote feeds into ClustalW Results Display. A text box at the bottom contains the instruction: "Drag and drop utilities, director and actor libraries."

Distributed Workflows in KEPLER

- Web and Grid Service plug-ins
 - WSDL (now) and Grid services (stay tuned ...)
 - ProxyInit, GlobusGridJob, GridFTP, DataAccessWizard
 - SSH, SCP, SDSC SRB, OGS?-???... *coming*
- WS Harvester
 - Import query-defined WS operations as Kepler actors
- XSLT and XQuery Data Transformers
 - to link **not** “designed-to-fit” web services

Generic Web Service Actor



- Given a WSDL and the name of an operation of a web service, dynamically customizes itself to implement and execute that method.

Configure - select service operation

Web Service Harvester (Ilkay Altintas, SDM)

file:/C:/Documents and Settings/altintas/Desktop/harvesterModel.xml

File View Edit Graph Debug Help

utilities
director library
actor library
sources
kepler
web services
BlastDemo_execute
Blast_searchParam
Blast_searchSimple
Blast_extractPosition
Blast_searchParamAsyt
Blast_searchSimpleAsyt
ClustaW_analyzeParam
ClustaW_analyzeSimple
ClustaW_analyzeParam
ClustaW_analyzeSimple
DDBJ_getFFEntry
DDBJ_getXMLEntry
DDBJ_getFeatureInfo
DDBJ_getAllFeatures
DDBJ_getRelatedFeatur
DDBJ_getRelatedFeatur
ExClustaW_analyzePar
ExClustaW_analyzeSimj
ExClustaW_analyzePar
ExClustaW_analyzeSimj
Fasta_searchParam

PN Director

WSHarvester

Edit parameters for WSHarvester

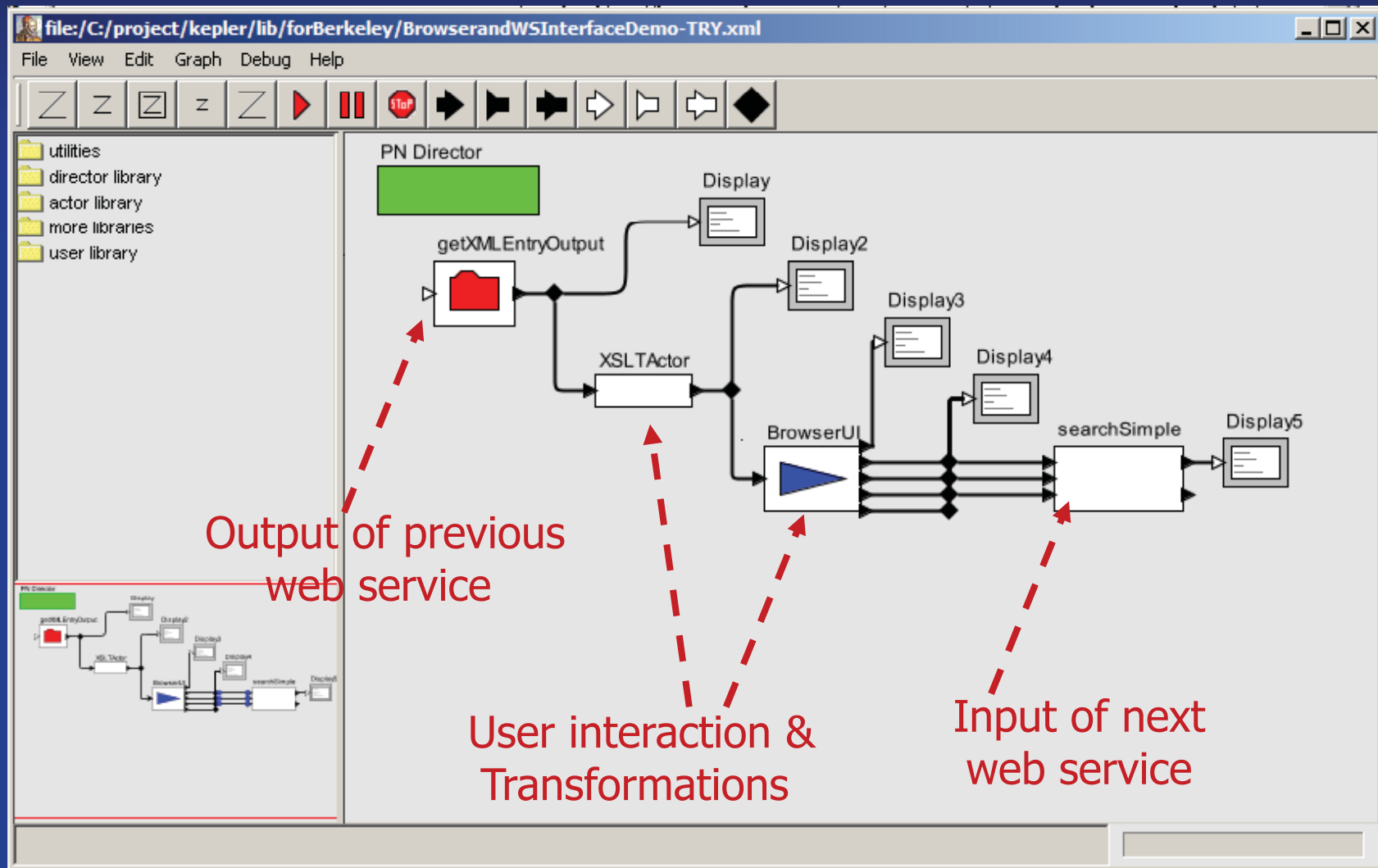
repositoryURL: Browse

keywords:

Commit Add Remove Edit Styles Help Cancel

- Imports the web services in a repository into the actor library.
- Has the capability to search for web services based on a keyword.

Composing 3rd-Party WSs (NMI, Steve Mock)



Framework Components

Archivists' Workbench
Archival Processes as Web Services

Portal Technology

Workflow Systems

■ **Data Grids & Federation**

IP2: General Studies

■ FOCUS 2

Persistent Archives Based on Data Grids

This study focuses on the San Diego Supercomputer Centre's project to develop a prototype for a persistent archive based upon data grid technology for the National Archives and Records Administration (NARA). The general study team will examine the minimal capabilities needed within grid technology for preservation of governmental records, focusing on activities related to the preservation of NARA's selected digital holdings.