Archivists' Toolkit: Ingest Functional Area

Outline

Ingest Functional Area

I1: Description

The Archivists' Toolkit will undoubtedly be deployed in repositories that have already invested considerable effort in producing EAD finding aids, as well as capturing and storing electronically information not necessarily contained in the encoded finding aids. So that repositories are able to integrate their archival information and not be required to manage and operate legacy information systems, it is extremely important that the Toolkit provide for automated ingest of information.

Ideally, the AT application will support ingesting 1) standard data such as EAD finding aids, EAC records, and MARC 21 records and 2) non-standard data such as accession information, source information, and location information stored in local electronic databases.

The objective of the first phase of the AT project will be to build functionality to support the ingest of legacy EAD finding aids and MARC 21 records. It is assumed that MARC records will only be ingested in those cases in which there is not a corresponding EAD encoded finding aid. There is no reason to ingest both an EAD document and a MARC record for the same archival resource, and in almost all cases, the EAD record for a resource will be a fuller and more detailed representation than will be the MARC record.

The rest of the Ingest specification is constructed according to this limitation. Subsequent phases of the AT project may address the need for functionality to support the ingest of other standardized data formats, such as EAC records should they proliferate and become generally available, or METS expressed digital objects, or local information stored electronically.

Data Supported by Ingest

Only ingest of EAD and MARC 21 records will be supported in the first version of the Toolkit. Both kinds of data need to be expressed as XML in order to be ingested.

The EAD must be a valid instance of version 1.0 or 2002. It will only map to and be stored as resource and resource component records. Successful ingest of an EAD shall be demonstrated by parsing of the EAD elements across the resource and resource component records and the ability to produce from the AT an EAD instance more or less identical to that ingested.

The MARC 21 must be valid also and be a parent record, i.e., not a child record to another MARC record. The MARC record will only map to and be stored as a resource record. It shall not be possible to ingest MARC child records or to map MARC records to resource component records. Successful ingest of a MARC 21 record shall be demonstrated by parsing of the MARC 21 descriptive data into an AT resource record and the ability to output that information as a “partial MARC record.”

For either EAD or MARC records, access points such as the controlled access values in EAD or the 1xx / 6xx / 7xx values in MARC shall be mapped to the appropriate name records or subject records. And either type of source record shall designate a resource ID, which must be unique in the context of a local implementation of the AT. Thus, it shall not be possible to ingest multiple MARC records for the same resource ID.

In either case of EAD or MARC records, audit information shall be added to the records created as a result of the ingest process. The audit information shall be set for the operator initiating the ingest command and include the date the ingested records were created.

Ingest Process

The operator will select the command to ingest resource records.

The operator should have the option to use a GUI interface to ingest single records. The operator should be able to browse files in a directory, select a file to be ingested, and then start the ingest process using a command button. (Ideally, it would be best to support ingest of record batches. To do so, it would probably be necessary to have in one batch (directory) only records of the same type, i.e., only records of EAD version 1.0 or version 2002 AND to have a reliable means for discerning the resource ID of the record to be ingested in order to check it for duplication against the implemented AT. While the EAD header may provide a reliable means for discerning the resource ID, it is not clear that the same would be true for MARC records.)

After identifying a record for ingest and selecting the ingest command, the operator will be asked to provide two pieces of data. One will be to indicate that the source record is an EAD record or a MARC record. The second will be to indicate the resource ID for the record to be ingested.

The machine will apply the ingest routine for the type of record to be ingested as expressed by the operator. If there is mismatch between record type expressed and record type in the file, the machine shall respond that the record can not be ingested because it is not of the type indicated by the operator. Otherwise the machine moves to assess the next piece of data.

If there is not a file mismatch, the machine will then check to assure that the resource ID expressed by the operator is not a resource ID already used in the toolkit. If the resource ID is already used in the local AT application, then the machine shall notify the operator that the record can not be ingested because another resource with that ID already exists in the local implementation of the AT. The operator shall be prompted to enter a different resource ID or to cancel the ingest process.

If the identity check for record type and resource ID are successful, the ingest process will begin.

First, the resource will be evaluated to assure that it is a valid EAD 1.0 or 2002 instance or a valid MARC 21 record. If it is not a valid record, the machine will indicate that the record is not valid and cannot be ingested. The machine will also indicate why the validation check failed. (In the case of a MARC record, another reason for a validation error is that the record is a child record and, as such, can not be ingested into the AT.) The operator shall have the option to print the validation error report or save it to a text file named by the operator. Finally, the operator will be prompted to ingest another record or cancel the ingest process.

If the record is a valid record, then the machine shall indicate that the record is valid and is being ingested into the AT implementation. This prompt shall remain on the monitor until the ingest process is completed. At that point, the machine shall indicate that the ingest process has successfully concluded The machine shall provide a display of all the AT records created as a result of the ingest process. The resource records shall be listed in hierarchical order with AT record numbers. The operator shall have the option to print the ingest results or save them to a text file named by the operator.

After the operator chooses to either save the results file or not, the machine shall prompt the operator to add the newly created records to the index of records or wait until after other additional records are ingested before indexing the database. If the operator chooses to wait, the machine will prompt the operator to select another record to be ingested. At this point, the ingest process will start over again from the beginning. If the operator chooses, to index the database, the machine will indicate that the database is being indexed. When the index is completed, the machine shall indicate the database has been successfully indexed.

I2: Business Rules:

1. Legacy records can be integrated into an AT implementation using the ingest function.

2. Legacy source records must be XML documents of either EAD version 1.0 or version 2002 or they must be MARC 21 records.

3. The legacy source record to be ingested must be represented by a resource ID not already in use in the AT implementation.

4. The legacy record to be ingested must be a valid instance of its type, i.e., it must be a valid EAD 2002 record (and, in the case of a MARC record, must not be a child record)

5. A valid EAD record will be ingested as resource and resource component records, and its controlled access values will be ingested appropriately as either subject or name records.

6. A valid MARC 21 record will be ingested as a resource record, and its 1xx/6xx/7xx values will be ingested appropriately as either subject or name records.

7. Documentation will be provided for all legacy records that cannot be ingested. Such documentation will indicate the reason that the validation check failed.

8. Documentation will be provided for all legacy records that are ingested. Such documentation will consist of the record IDs and record structure (in the case of EADs).

I3: Required Tasks

1. Operator selects option to ingest legacy records

2. Operator selects legacy record to be ingested

3. Operator indicates if legacy record is EAD version 1.0, EAD version 2.0, or MARC 21 record

4. Operator indicates resource ID for resource represented by legacy record

I4: Optional Tasks

1. Operator chooses to save validation error report to file

2. Operator chooses to save ingest report to file

3. Operator chooses to cancel ingest process prior to ingest successfully completing

4. Operator chooses to ingest additional legacy records

I5: User Input / Machine Response

User Intentions *(Required in Italics)*	Application Response / Action
Operator selects option to ingest a legacy record	
	Machine returns a window with browse function for scanning directories and identifying the file containing the legacy record The browse window also includes two additional frames. One is for indicating if the file contains a record of the type: 1) EAD version 1.0, 2) EAD version 2002, or 3) MARC 21 The other is for indicating the resource number to be used by the resource represented in the legacy record.
Operator uses browse window to specify the file containing the legacy record. Operator indicates that the file type contains: 1) EAD version 1.0, 2) EAD version 2002, or 3) MARC 21. Operator indicates the resource number to be used for the ingested record. Operator selects command to ingest the record contained in the file.	
	Machine displays message that “Record is being validated”
	If there is a mismatch between record type expressed in the query window

	and the record type contained in the file, validation process will halt and machine will display message: “Record can not be ingested. Not the same record type specified by operator.” Otherwise, the machine proceeds with the validation check, continuing to display the message that “Record is being validated”.
	If the resource ID entered by the operator is already used in the local implementation of the AT, the machine will halt the ingest process and display the message “Record can not be ingested. Resource ID is already used in local database. Resource ID must be a unique number.” Otherwise, the machine proceeds with the validation check, continuing to display the message that “Record is being validated.”
	If the legacy record is a MARC record and is coded as a child record, then the machine stops the validation process, displaying the message “Record can not be ingested. MARC record is a child record. Child records can not be ingested in to the AT.” Otherwise, the machine proceeds with the validation check, continuing to display the message that “Record is being validated.”
	If either the EAD xml file or MARC xml file are not valid EAD or MARC xml files, the machine halts the validation process and displays the message “Record can not be ingested. File is not valid. See details below and try to ingest file again after correcting the

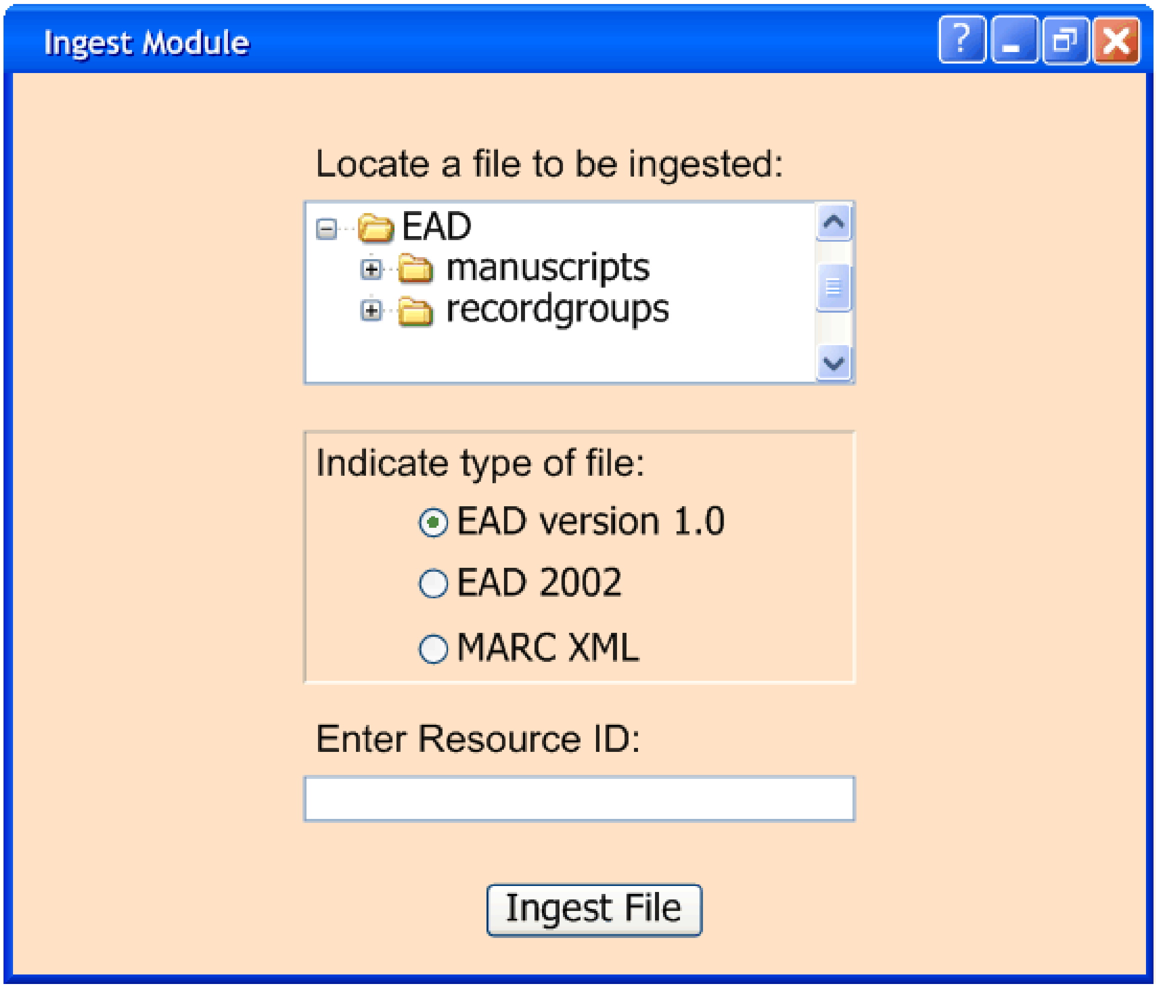
	error.” This message is followed by an error report, which should describe the characteristics of the error and locate it using a line number index keyed to the source file. The display should have command options for printing the error report directly or saving it to a file.
	Following printing / saving of the error report, the machine asks the operator is another record should be ingested.
Operator selects the option “No”	
	Operator is returned to place where command to ingest a legacy record was first selected, i.e., main menu
Operator selects the option “Yes”	
	Operator is returned to record identification window, where operator selects another record, indicating its type and resource ID.
	Otherwise, the machine continues validating the record, displaying the message “Record is being validated.”
	Once the validation process is completed without a validation error, machine reports and asks: “Record is valid. Do you want to ingest it now?”
Operator selects “No”	
	Machine responds “Are you sure you do not want to ingest the validated record now?”
Operator selects “Yes”	
	Operator is returned to main menu, where command to ingest a legacy record was first selected.
Or the operator indicates the record is to be ingested on either the first query or the second query.	
	Machine responds with display “Record is being ingested. Please wait.”

	When the ingest process is complete, the machine responds “Record for [Resource ID] has been ingested.” Following that statement is a list identifying the records created as a result of the ingest process. With the list are command options for printing the ingest report or saving it to a text file.
	Following printing / saving of the ingest report, the machine asks the operator if another record is to be ingested.
Operator selects the option “No”	
	Operator is returned to place where command to ingest a legacy record was first selected, i.e., main menu
Operator selects the option “Yes”	
	Operator is returned to record identification window, where operator selects another record, indicating its type and resource ID.

I6: Reports

- Ingest Validation Error Report
- Ingest Success Report

I7: Screen Sequences



Ingest Module

Locate a file to be ingested:

- EAD
  - manuscripts
  - recordgroups

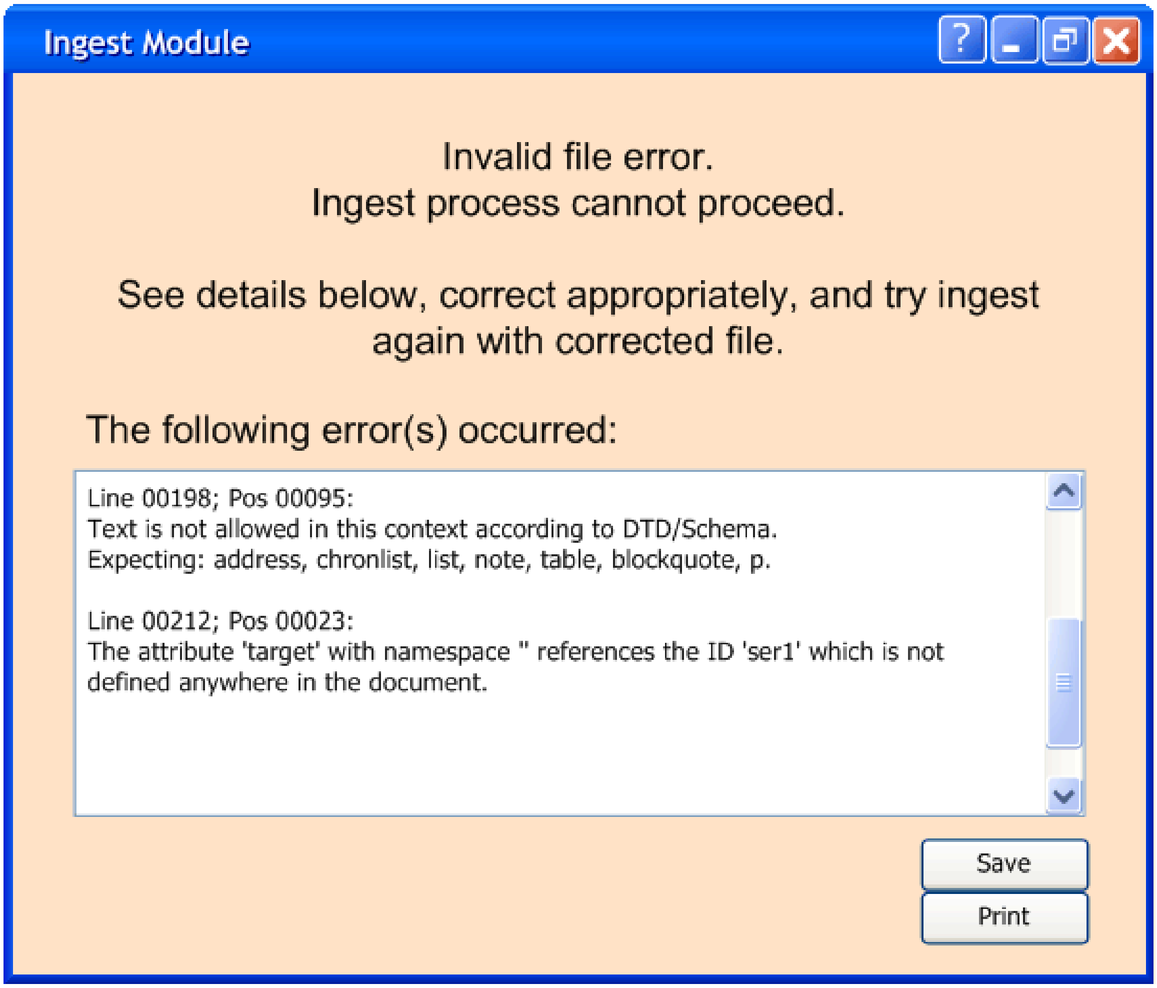
Indicate type of file:

● EAD version 1.0

○ EAD 2002

○ MARC XML

Enter Resource ID:



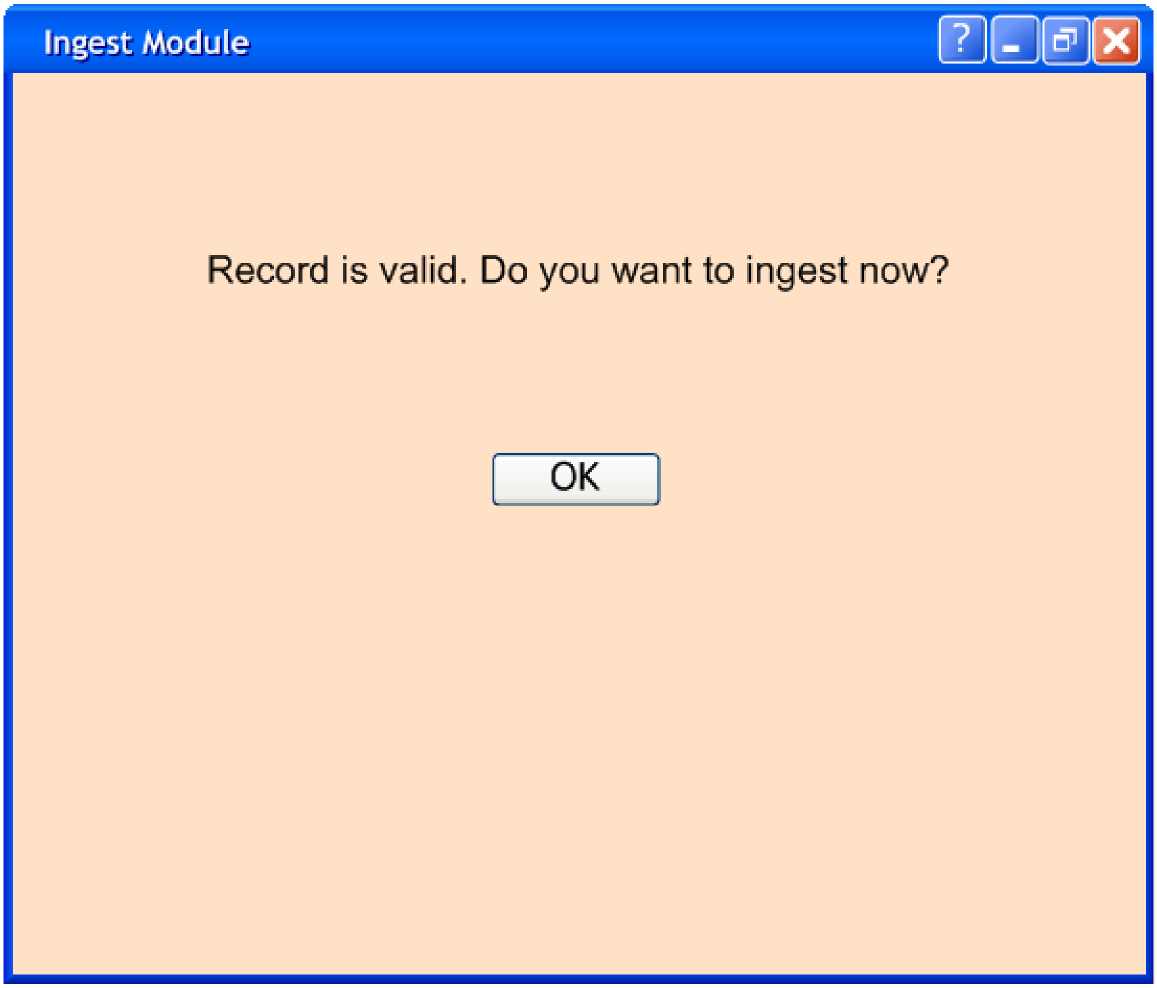
Invalid file error.
Ingest process cannot proceed.

See details below, correct appropriately, and try ingest again with corrected file.

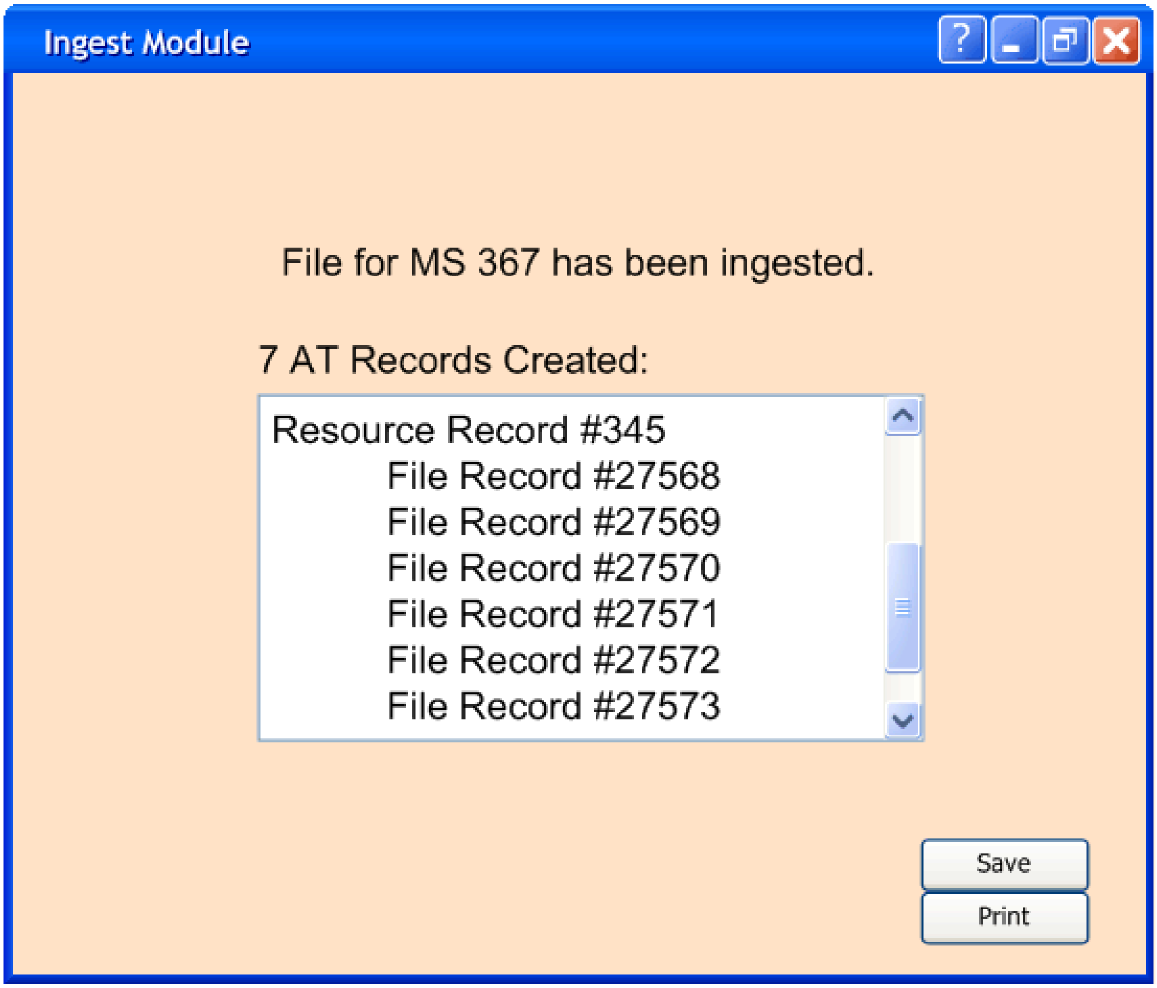
The following error(s) occurred:

```
Line 00198; Pos 00095:
Text is not allowed in this context according to DTD/Schema.
Expecting: address, chronlist, list, note, table, blockquote, p.

Line 00212; Pos 00023:
The attribute 'target' with namespace '' references the ID 'ser1' which is not
defined anywhere in the document.
```



Record is valid. Do you want to ingest now?



File for MS 367 has been ingested.

7 AT Records Created:

Resource Record #345
- File Record #27568
- File Record #27569
- File Record #27570
- File Record #27571
- File Record #27572
- File Record #27573