MAGPIE

Metadata Assignment GUI Providing Ingest and Export

Stephanie Elmquist, William Welling (corresponding author), James Creel, Jeremy Huff, Jason Savell, Rincy Mathew, Doug Hahn, and Michael Bolton
Magpie

- **Magpies** are **birds** of the **Corvidae** (crow) family, including the black and white **Eurasian magpie**, which is considered one of the most intelligent animals in the world, and the only non-mammal species able to recognize itself in a **mirror test**.[1] In addition to other members of the genus **Pica**, corvids considered as magpies are in the genera **Cissa**, **Cyanopica** and **Urocissa**.

- **MAGPIE** is an acronym for “Metadata Assignment GUI Providing Ingest and Export”

Objectives

• Intuitive and efficient interface for managing metadata annotation of digital content
• Consume external resources to provide definitive records and suggestions
• Providing single and batch export into Institutional Repository (IR), i.e. DSpace
• Generation and export of archival cataloging spreadsheets for preservation
Use Cases

• Legacy dissertations
• Image collections
• Newspapers and articles
• Etc.
Implementation

• Using A&M developed Weaver framework
• Distributed architecture
  – Scanning workstation
  – ABBYY OCR
  – Authorization Service *Weaver
  – Webservice *Weaver
  – User Interface *Weaver
  – Institutional Repository
  – Archivematica

* Weaver Webservice Core  * Weaver UI Core
Project Configuration

• External JSON for metadata input fields
  – label, gloss, repeatable, and input type
• Network Directory Watcher (Polling)
  – New project directories
  – Project folder for scanned documents
  – Map folder for DSpace ingest map files
• External Services
  – Voyager for MARC records
Workflow

- Provide Digital Content to Watcher
  - For example, scan/ocr print document
- Automatic annotation from definitive sources
- Assign, Annotate, Review, and Approve annotations
- Export for ingest into IR or directly ingest
- Export preservation spreadsheet
Workflow Diagram

Metadata Tool: Dissertation Project Workflow

1. Scan Dissertation
2. Copy Image To OCR Folder
3. Pre-Populate with MARC
4. Assign Annotator
5. Annotate
6. Approve Metadata
7. Export Metadata
8. Perform Ingest
9. Return Map File
10. Update Status (unlocking)
11. Export Metadata CSV
12. Quality Assurance
13. Build Package
14. Archive Package

Directory on Network Drive
Web App
Web Service API
Document Image, OCR Text, and Metadata CSV
Data Center

Manager / Metadata Librarian
Annotator
Annotator
Annotator
Dsparse Admin
Preservation Librarian

Repository
ABBY OCR
Voyager (OPAC)

Scanned Document (tif)
OCR Folder (Monitored)
Dissertation Folder (Watcher Service)
Preservation Folder
Export Folder
Map File Folder (Watcher Service)

Dsparse
Archivematica (DPN)
Metadata

• Labels
  – Authors
  – Subjects
  – Abstract
  – Publication dates
  – Etc.

• Justification
  – Discovery
    • Indexing
    • Categorization
  – Preservation
Benefits for IR Content

• DSpace is the first IR to which MAGPIE exports
  – via CSV (with supporting tools), SAF, or REST API
• Facilitates controlled vocabulary
• Normalizes metadata
• Allows opportunity for curation of new and legacy metadata
• Custom forms for specific projects
Brief Demonstration

- User management
- Document management
  - Assign
  - Annotate
  - Review
  - Publish
- Export
  - CSV
  - SAF
Future Direction

• Semi-automatic abstract, author, and citation extraction
• Statistical Classification
• Topic Modeling
Thank You.

Any Questions?

Stephanie Elmquist – s-elmquist@library.tamu.edu
William Welling – wwelling@library.tamu.edu
James Creel – jcreel@library.tamu.edu
Jeremy Huff – huff@library.tamu.edu
Jason Savell – jsavell@library.tamu.edu
Rincy Mathew – rmathew@library.tamu.edu
Doug Hahn – dhahn@library.tamu.edu
Michael Bolton – michael.bolton@library.tamu.edu