Introduction

- Digital Repository
  - Initial focus on digitized book and journal content
  - “Light” archive
- Collections and Collaboration
  - Comprehensive collection
  - Shared strategies
  - Local services
  - Public Good

Delivering Data For New Generations of Research

Strategies and Challenges

Jeremy York
NISO/BISG Forum
ALA Annual 2010

Content Distribution

- 6,173,575 – Total
- 1,177,667 – Public Domain

* As of June 15, 2010

Language Distribution (1)

The top 10 languages make up ~86% of all content

- English 48%
- German 8%
- French 7%
- Russian 5%
- Spanish 4%
- Chinese 4%
- Japanese 4%
- Italian 3%
- Arabic 2%
- Polish 1%

* As of June 15, 2010

Language Distribution (2)

The next 40 languages make up ~13% of total

* As of June 15, 2010

Originating Institution

* As of June 15, 2010
Content over time

Content Growth

Data Distribution & APIs
- OAI-PMH
- Metadata files
- Bibliographic API
- Data API

Extended Services
- Community Development Environment
- Non-Google Ingest
- Non-Book/Non-Journal Ingest
- Computational Research

Strategies for Computational Research
- Data distribution
- Protocol-based access
- Research Center
SEASR Architecture

- Visualizations
  - User interfaces
  - Repositories
  - Data Analysis Components
  - Components
  - Cloud Computing

SEASR @ Work – Tag Cloud
- Count tokens
- Filter options supported
- Stem words

SEASR @ Work – Entity Mash-up
- Entity Extraction with OpenNLP or Stanford NER
- Locations viewed on Google Map
- Dates viewed on Simile Timeline

SEASR @ Work – Entities To Network
- Identify entities
- Define relationships between entities within same sentence

SEASR @ Work – Text Clustering
- Clustering of Text by token counts
- Filtering options for stop words, Part of Speech
- Dendogram Visualization
SEASR @ Work – Audio Analysis

- **NEMA**: Executes a SEASR flow for each run
  - Loads audio data
  - Extracts features for every 10 sec moving window of audio
  - Loads and applies the models
  - Sends results back to the WebUI
- **NESTER**: Annotation of Audio via Spectral Analysis

SEASR @ Work – Zotero

- Plugin to Firefox
- Zotero manages the collection
- Launch SEASR Analytics
  - Citation Analysis uses the JUNG network importance algorithms to rank the authors in the citation network that is exported as RDF data from Zotero to SEASR
  - Zotero Export to Fedora through SEASR
  - Saves results from SEASR Analytics to a Collection
- Launch MONK Processing
  - MONK DB Ingestion Workflow

SEASR @ Work – Emotion Tracking

Goal is to have this type of Visualization to track emotions across a text document (Leveraging flare.prefuse.org)

Sentiment Analysis: Visualization

Person Extraction:
Scott’s *Waverley*, *Ivanhoe*, and The *Heart of Midlothian*.

Location Extraction:
Top: Walter Scott’s *Waverley*  
Bottom: Maria Edgeworth’s *Castle Rackrent*
Thank you!

hathitrust-info@umich.edu
jjyork@umich.edu