



Delivering Data For New Generations of Research

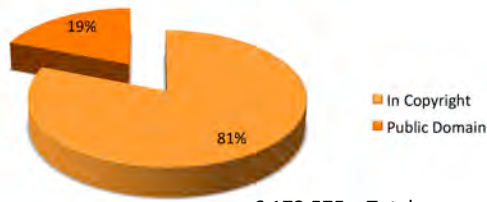
Strategies and Challenges

Jeremy York
NISO/BISG Forum
ALA Annual 2010

Introduction

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

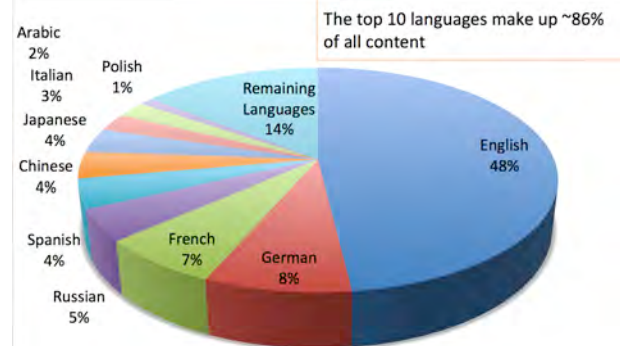
Content Distribution



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

* As of June 15, 2010

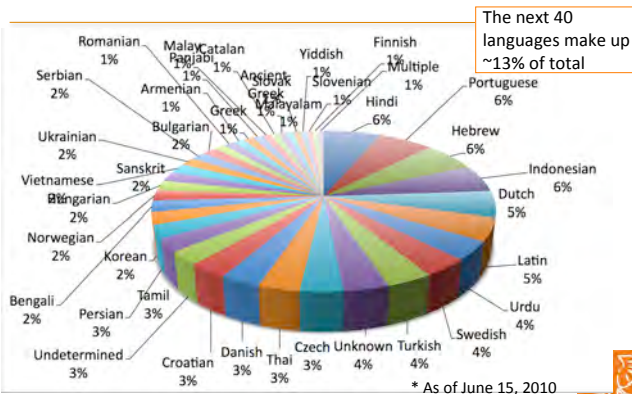
Language Distribution (1)



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

* As of June 15, 2010

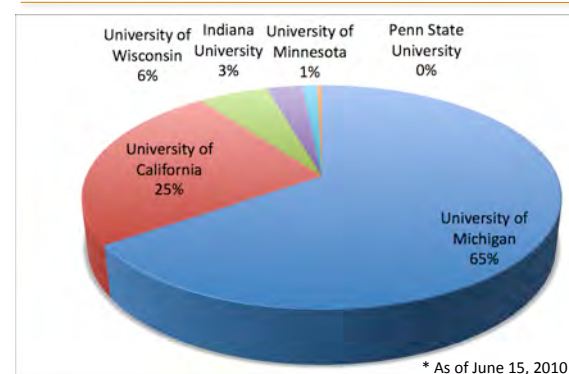
Language Distribution (2)



The next 40 languages make up ~13% of total

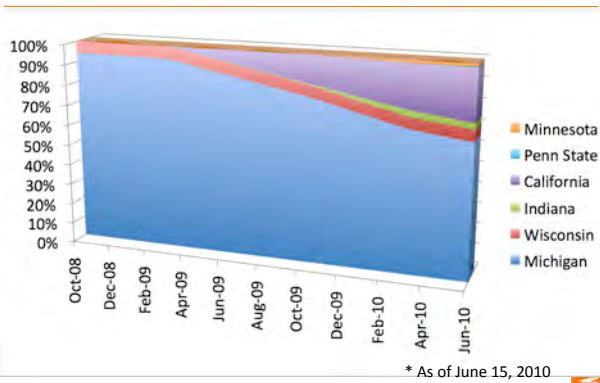
* As of June 15, 2010

Originating Institution

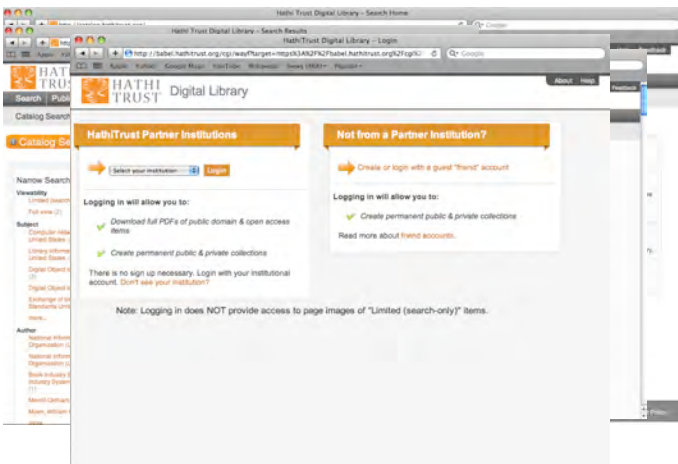
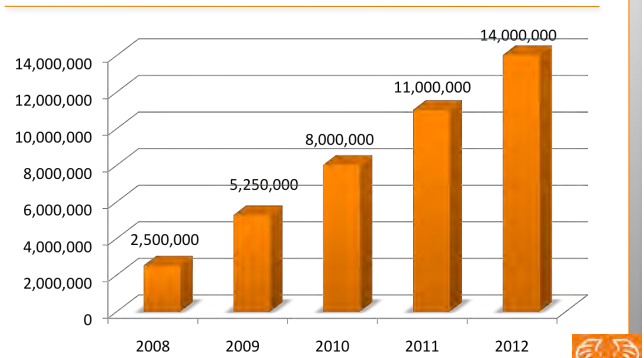


* As of June 15, 2010

Content over time



Content Growth



Data Distribution & APIs

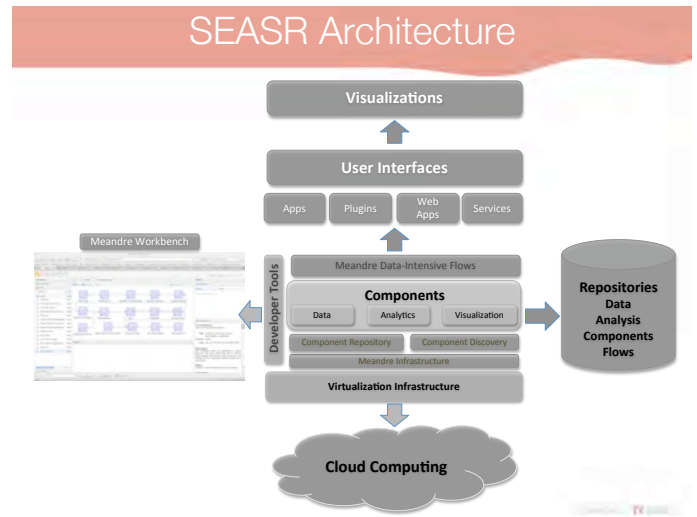
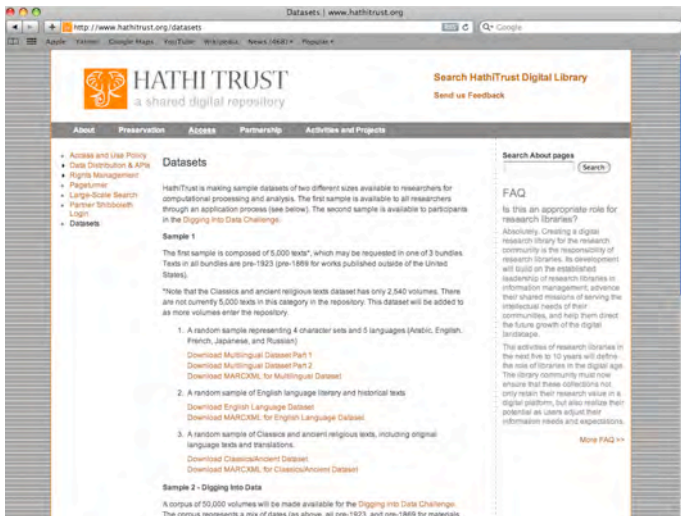
- OAI-PMH
- Metadata files
- Bibliographic API
- Data API

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



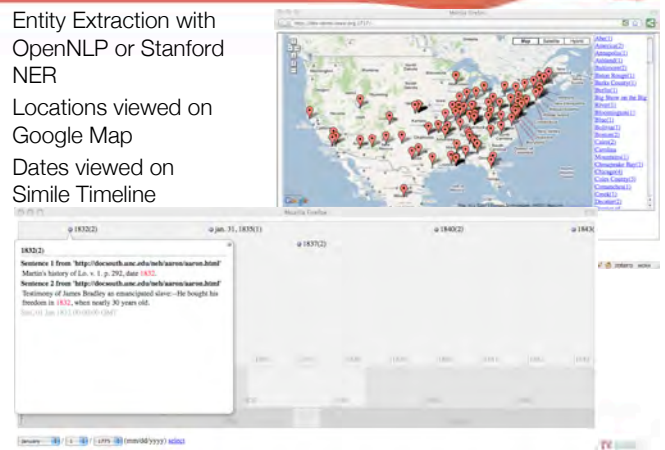
SEASR @ Work – Tag Cloud

- Count tokens
- Filter options supported
- Stem words



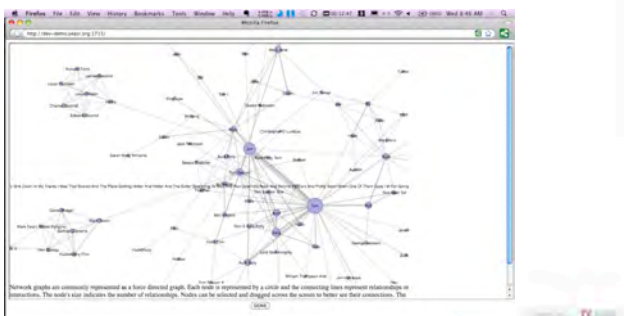
SEASR @ Work – Entity Mash-up

- Entity Extraction with OpenNLP or Stanford NER
- Locations viewed on Google Map
- Dates viewed on Simile Timeline



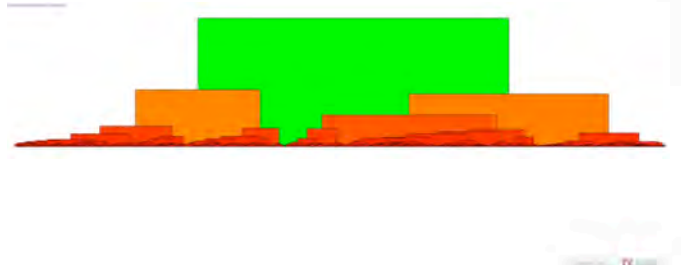
SEASR @ Work – Entities To Network

- 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
- Dendrogram Visualization



Thank you!

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

