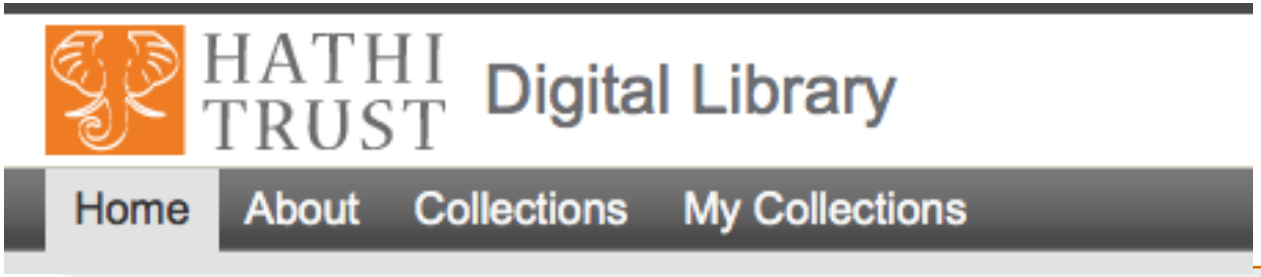
Digital Humanities At Scale: Hathi Trust Research Center

Beth Plale

Co-Director, HathiTrust Research Center

Professor, School of Informatics and Computing

Indiana University



Currently Digitized

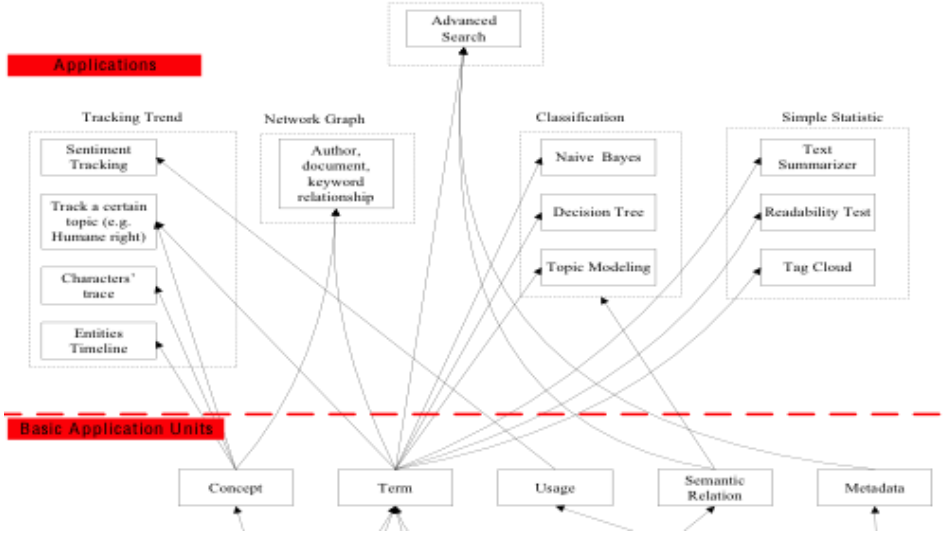
10,100,278 total volumes
5,345,001 book titles
266,113 serial titles
3,535,097,300 pages
453 terabytes
120 miles
8,206 tons
2,784,331 volumes (~28% of total) in the public domain
View visualizations of HathiTrust call numbers, languages, and dates

statistics information ››

→ HathiTrust Research Center: dedicated to provision of computational access to comprehensive body of published works for scholarship and education

→Big data and copyrighted text means ***computation moves to data, not vice versa***

HathiTrust Research Center

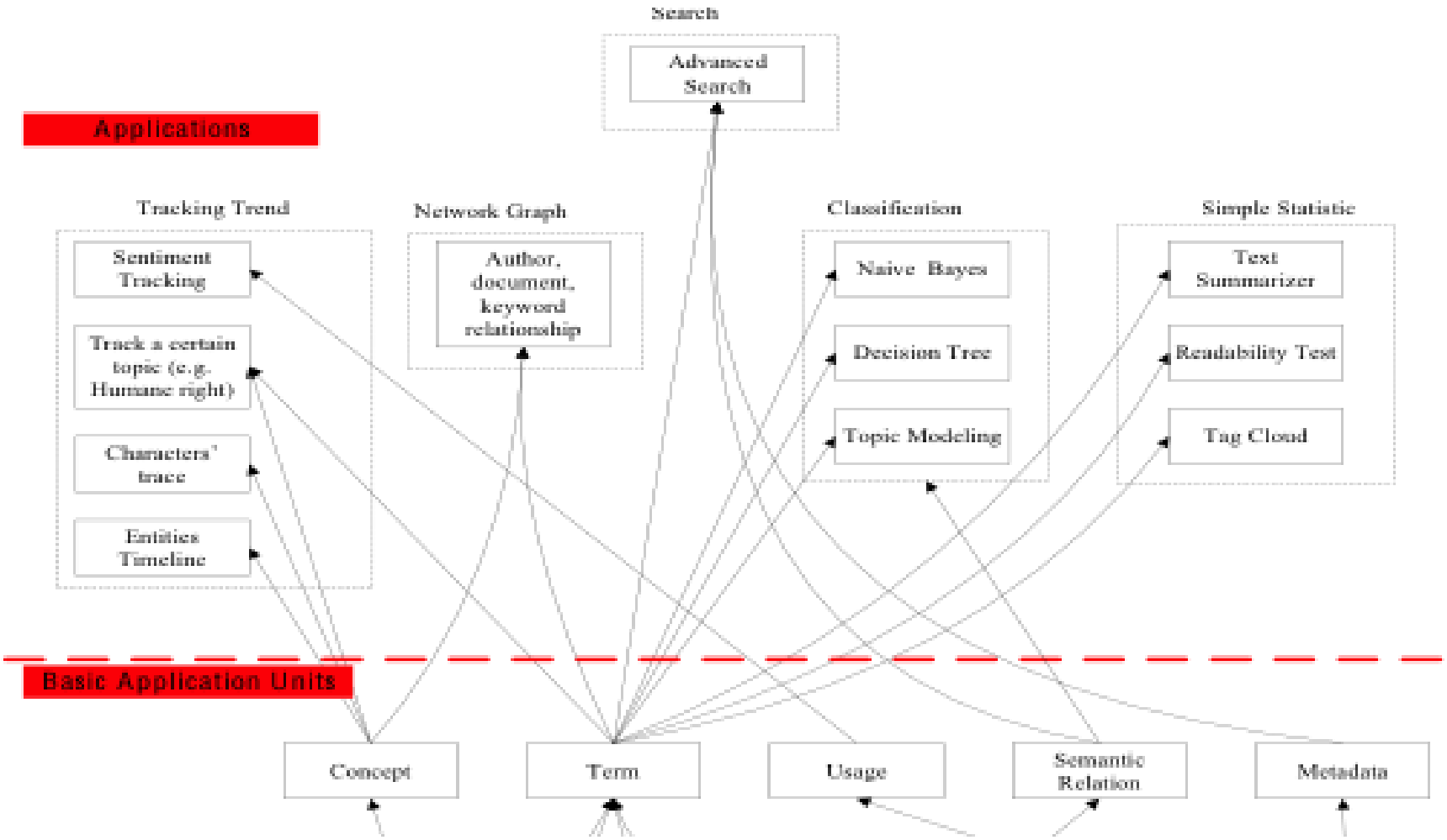


- Analysis on 10,000,000+ volumes of HathiTrust digital repository
- Working with OCR
- Large-scale data storage and access
- HPC and Cloud

Type of Data (Public domain and copyrighted works)	Estimated initial size: 300-500 TB
Solr Indexes	36 TB (3 indexes)
File system rsync	12 TB
Fast volume access store	30TB
Versions of collection (5)	120 TB
Volume store indexes	100 TB

Query = author:Withers, Hartley

Categories of algorithms



Parting thoughts

- UnCamp Sept 20-21, 2012 Bloomington, IN
- Currently setting up support for public domain corpus
- Scholarly community interest in this resource is high; it's value proposition is clear