
Microsoft® Research
Faculty Summit
10
YEAR ANNIVERSARY

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

The iSchool AT ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Text-Mining and Humanities Research

Microsoft Faculty Summit, July 2009

John Unsworth

Topics:

- Why text-mining?

- What kinds of research questions can humanities scholars address with text-mining tools, and who is doing this kind of work?

- What kind of work needs to be done to prepare text collections for this kind of work, and what challenges face those who want to build text-mining software for this audience?

- What's next? And who funds it?

Why Text-Mining?

"The Greek historian Herodotus has the Athenian sage Solon estimate the lifetime of a human being at c. 26,250 days (Herodotus, The Histories, 1.32). If we could read a book on each of those days, it would take almost forty lifetimes to work through every volume in a single million book library. . . .

While libraries that contain more than one million items are not unusual, print libraries never possessed a million books of use to any one reader."

-- Greg Crane, "What Do You Do With A Million Books?" D-Lib Magazine, March 2006, Volume 12 Number 3

Why Text-Mining?

"Ten years ago ... a young Jesuit named Roberto Busa at Rome's Gregorian University chose an extraordinary project for his doctor's thesis in theology: sorting out the different shades of meaning of every word used by St. Thomas Aquinas. But when he found that Aquinas had written 13 million words, Busa sadly settled for an analysis of only one word—the various meanings assigned by St. Thomas to the preposition "in." Even this took him four years, and it irked him that the original task remained undone ... But in seven years IBM technicians in the U.S. and in Italy, working with Busa, devised a way to do the job. The complete works of Aquinas will be typed onto punch cards; the machines will then work through the words and produce a systematic index of every word St. Thomas used, together with the number of times it appears, where it appears, and the six words immediately preceding and following each appearance (to give the context). This will take the machines 8,125 hours; the same job would be likely to take one man a lifetime." --*Time*, December 31, 1956

Research Questions

DHQ: Digital Humanities Quarterly
Spring 2009
Volume 3 Number 2

Special Cluster: Data Mining
Editor: Mark Olsen

Words, Patterns and Documents: Experiments in Machine Learning and Text Analysis
Shlomo Argamon, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology; Mark Olsen, ARTFL Project, University of Chicago

Research Questions

Vive la Différence! Text Mining Gender Difference in French Literature

- 300 male-authored and 300 female-authored French texts classified for author gender using SVM, at 90% accuracy
- Results exhibit remarkable cross-linguistic parallels with results from a similar study of the British National Corpus
- Female authors use personal pronouns and negative polarity items at a much higher rate than their male counterparts
- Male authors demonstrate a strong preference for determiners and numerical quantifiers

RQ1: Vive la Différence!

"Among the words that characterize male or female writing consistently over the time period spanned by the corpus, a number of cohesive semantic groups are identified. Male authors, for example, use religious terminology rooted in the church, while female authors use secular language to discuss spirituality. Such differences would take an enormous human effort to discover by a close reading of such a large corpus, but once identified through text mining, they frame intriguing questions which scholars may address using traditional critical analysis methods."

Novels in English by Men, 1780-1900

m·o·n·k
metadata offer new knowledge

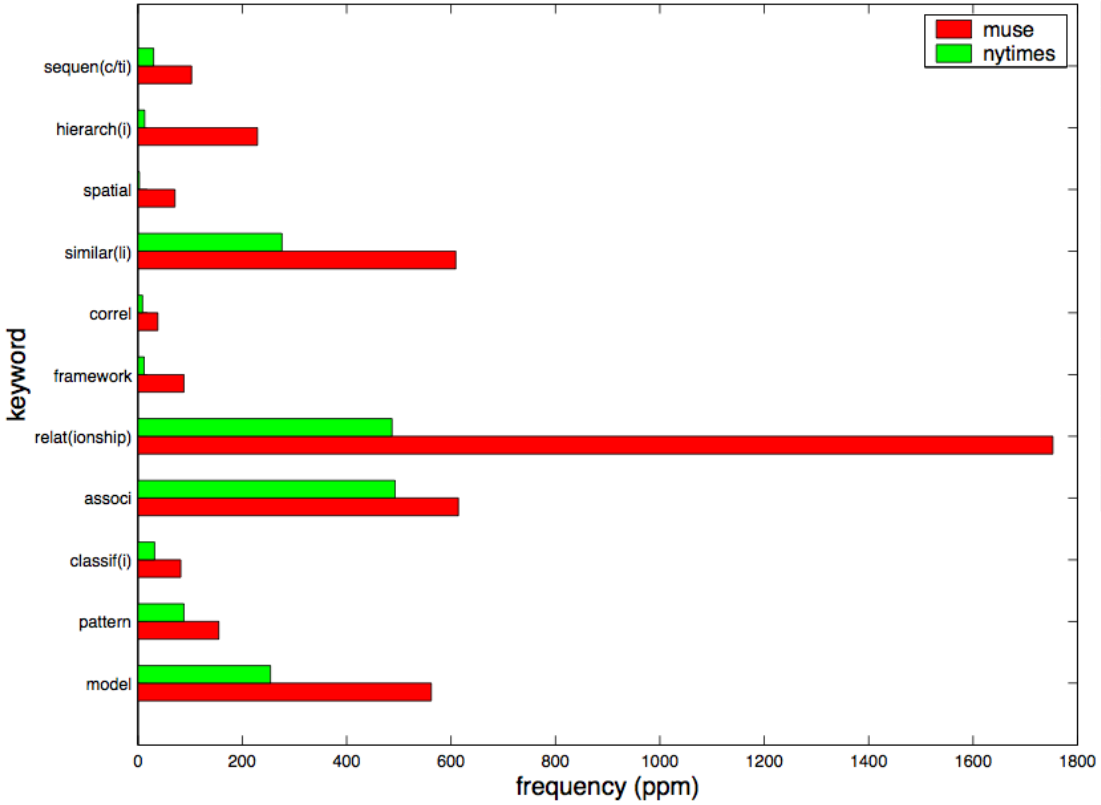
Analysis by Sara Steger and the MONK Project
Visualization by Wordle, from IBM's ManyEyes

Novels in English by Women, 1780-1900

m·o·n·k
metadata offer new knowledge

Analysis by Sara Steger and the MONK Project
Visualization by Wordle, from IBM's ManyEyes

MUSE Journals vs. New York Times



keyword	kdd-tf	muse-df:tf	muse-df-pcnt:tf-ppm	nytimes-df:tf	nytimes-df-pcnt:tf-ppm
cluster	52	13:17	10:22	50:56	1:24
model	40	99:438	**80:562**	335:597	**8:254**
pattern	29	49:121	**40:155**	167:207	**4:88**
network	23	30:76	24:97	325:724	8:307
classif classifi	35	14 19 :64	**11 15 :82**	16 50 :74	**0 1 :32**
rule	19	81:210	65:269	708:1409	17:598
associ	15	103:479	**83:614**	762:1161	**18:493**
graph graphic	15	2 15 :25	2 12 :32	7 244 :504	0 6 :214
stream	15	19:26	15:33	103:117	2:50
serial seri	10	20 80 :213	16 65 :273	32 537 :946	1 13 :403
relat relationship	10	117 98 :1367	**94 79 :1753**	386 323 :911	**9 8 :487**
framework	10	38:69	**31:88**	25:28	**1:12**
correl	9	20:30	**16:38**	15:21	**0:9**
similar similarli	9	99 66 :475	**53 80 :609**	484 77 :649	**12 2 :276**
spatial	7	19:55	**15:71**	8:8	**0:3**
decis	7	57:161	46:206	683:1110	16:471
hierarch hierarchi	6	20 41 :178	**16 33 :229**	4 4 :45	**0 0 :13**
sequenc sequenti	6	34 8 :80	**6 27 :103**	4 51 :71	**0 1 :30**

Table A.1: KDD keyword frequency comparison between MUSE and ANC-NYTIMES

Bei Yu, "An Evaluation of Text-Classification Methods For Literary Study" Dissertation, GSLIS, University of Illinois, Urbana-Champaign, 2006

Figure A.1: TF of Selected KDD keywords common in MUSE but not common in ANC-NYTIMES

Research Questions

- Gender, Race, and Nationality in Black Drama, 1950-2006: Mining Differences in Language Use in Authors and their Characters
- Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie
- Cultural Capital in the Digital Era: Mapping the Success of Thomas Pynchon
- Corpus Analysis and Literary History
- "The Story of *One* or, Reread The Making of Americans by Gertrude Stein"
- "More Than a Feeling: Patterns in Sentimentality in Victorian Literature"
- "The Devil and Mother Shipton: Serendipitous Associations and the MONK Project"

Challenges

- Text represents language, which changes over time (spelling)
- Comparison of texts as data requires some normalization (lemma)
- Counting as a means of comparison requires having units to count (tokenization)
- Treating texts as data will entail processing a new representation of the texts, in order to make the texts comparable and make their features countable.

C1 : Challenge | Chalange | Caleng | Challanss | Chalenge

“A word token is the spelling or surface of form of a word. MONK performs a variety of operations that supply each token with additional 'metadata'. Take something like 'hee louyd hir depely'. This comes to exist in the MONK textbase as something like

hee_pns31_he louyd_vvd_love hir_pno31_she depely_av-j_deep

Because the textbase 'knows' that the surface 'louyd' is the past tense of the verb 'love' the individual token can be seen as an instance of several types: the spelling, the part of speech, and the lemma or dictionary entry form of a word.”

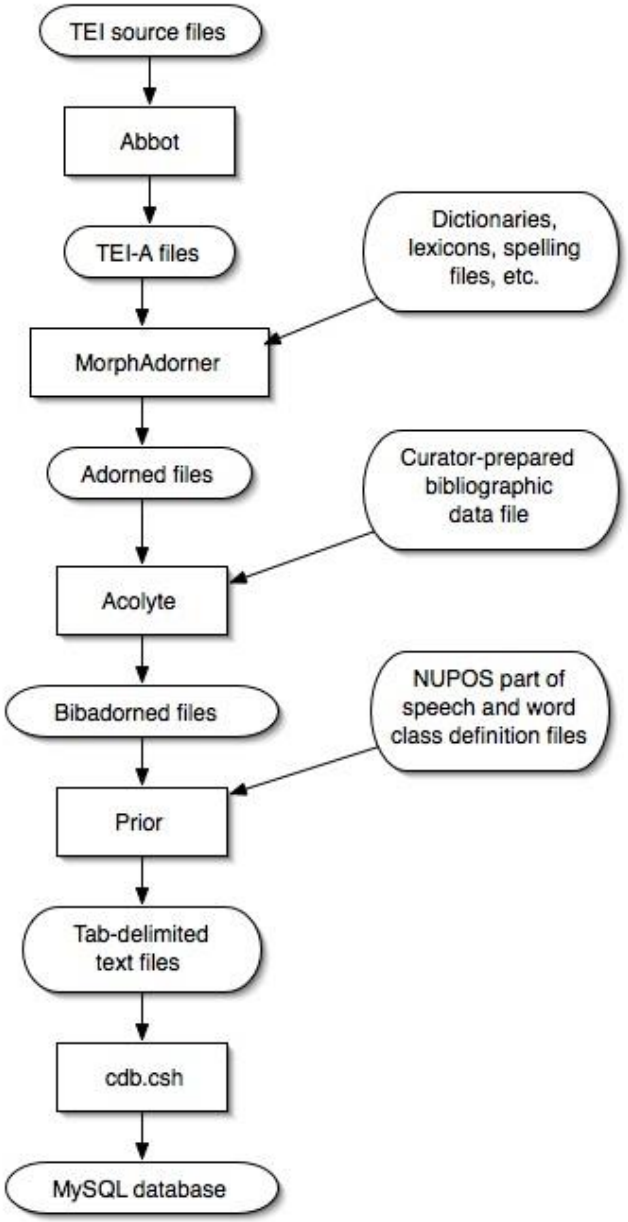
---Martin Mueller

C2: Reprocessing

MONK ingest process:

1. Tei source files (from various collections, with various idiosyncracies) go through Abbot, a series of xsl routines that transform the input format into TEI-Analytics (TEI-A for short), with some curatorial interaction.

2. “Unadorned” TEI-A files go through Morphadorner, a trainable part-of-speech tagger that tokenizes the texts into sentences, words and punctuation, assigns ids to the words and punctuation marks, and adorns the words with morphological tagging data (lemma, part of speech, and standard spelling).



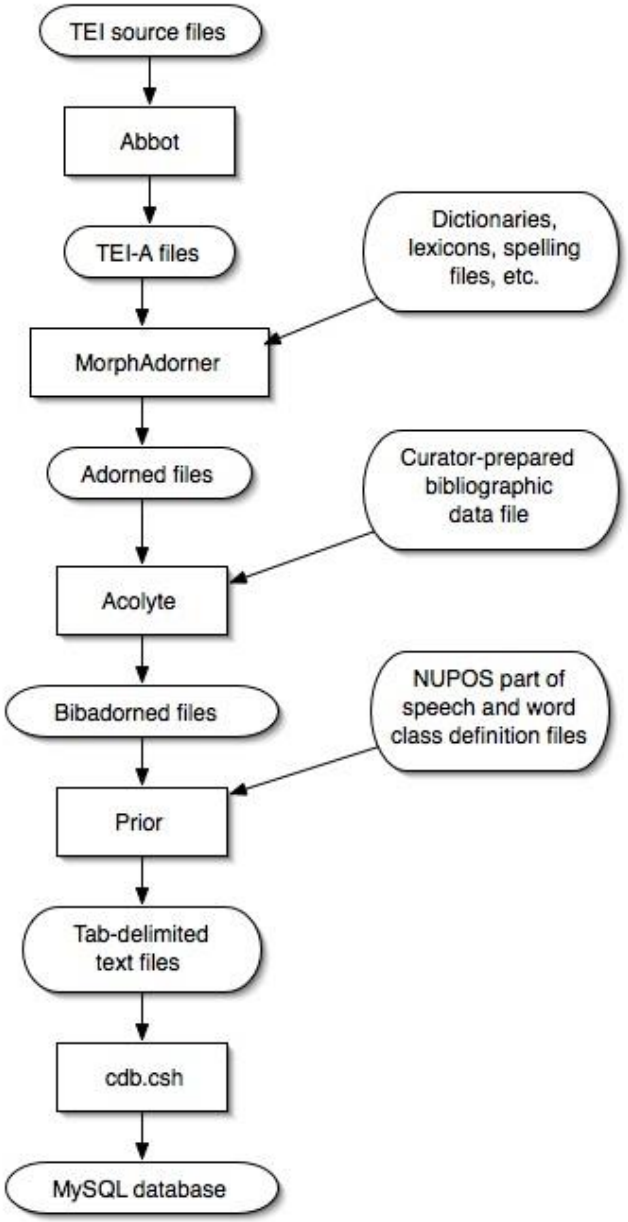
C2: Reprocessing

MONK ingest process (cont.):

3. Adorned TEI-A files go through Acolyte, a script that adds curator-prepared bibliographic data

4. Bibadorned files are processed by Prior, using a pair of files defining the parts of speech and word classes, to produce tab-delimited text files in MySQL import format, one file for each table in the MySQL database.

5. cdb.csh creates a Monk MySQL database and imports the tab-delimited text files.



C2: reprocessing

<docImprint>ENTERED according to Act of Congress, in the year 1867, by A. SIMPSON & CO.,<lb/>in the Clerk's Office of the District Court of the United States<lb/>for the Southern District of New York.</docImprint>

<docImprint>
<w eos="0" lem="enter" pos="vvn" reg="ENTERED" spe="ENTERED" tok="ENTERED" xml:id="allen-000600" ord="33" part="N">ENTERED</w>
<c> </c>
<w eos="0" lem="accord" pos="vvg" reg="according" spe="according" tok="according" xml:id="allen-000610" ord="34" part="N">according</w>
<c> </c>

Representation is 10X original (150MB becomes 1.5GB; 90% metadata); MONK is 150M words, but about 180 GB as a database, with indices, etc.

C2: Representation

"In the MONK project we used texts from TCP EEBO and ECCO, Wright American Fiction, Early American Fiction, and DocSouth -- all of them archives that proclaimed various degrees of adherence to the earlier [TEI] Guidelines.

Our overriding impression was that each of these archives made perfectly sensible decisions about this or that within its own domain, and none of them paid any attention to how its texts might be mixed and matched with other texts. That was reasonable ten years ago. But now we live in a world where you can multiple copies of all these archives on the hard drive of a single laptop, and people will want to mix and match."

--Martin Mueller

C2: Representa-tion

“Soft hyphens at the end of a line or page were the greatest sinners in terms of unnecessary variance across projects, and they caused no end of trouble. . . . The texts differed widely in what they did with EOL phenomena. The DocSouth people were the most consistent and intelligent: they moved the whole word to the previous line.... DocSouth texts also observe linebreaks but don't encode them explicitly. The EAF texts were better at that and encoded line breaks explicitly. The TCP texts were the worst: they didn't observe line breaks unless there was a soft hyphen or a missing hyphen, and then they had squirrelly solutions for them. The Wright archive used an odd procedure that, from the perspective of subsequent tokenization, would make the trailing word part a distinct token.”

-- Martin Mueller

C3: Features, Metadata, Interface

Tools can't operate on features unless those features are made available: for example,

- In order to count parts of speech (noun, verb, adjective) those parts have to have been identified.
- In order to find all the fiction by women in a collection, your data has to include information about genre and gender, and your interface has to allow you to select those facets.
- In order to find patterns, both the data and the interface have to support pattern-finding.
- Users like simple interfaces, but simple interfaces limit complex operations

What's Next? SEASR

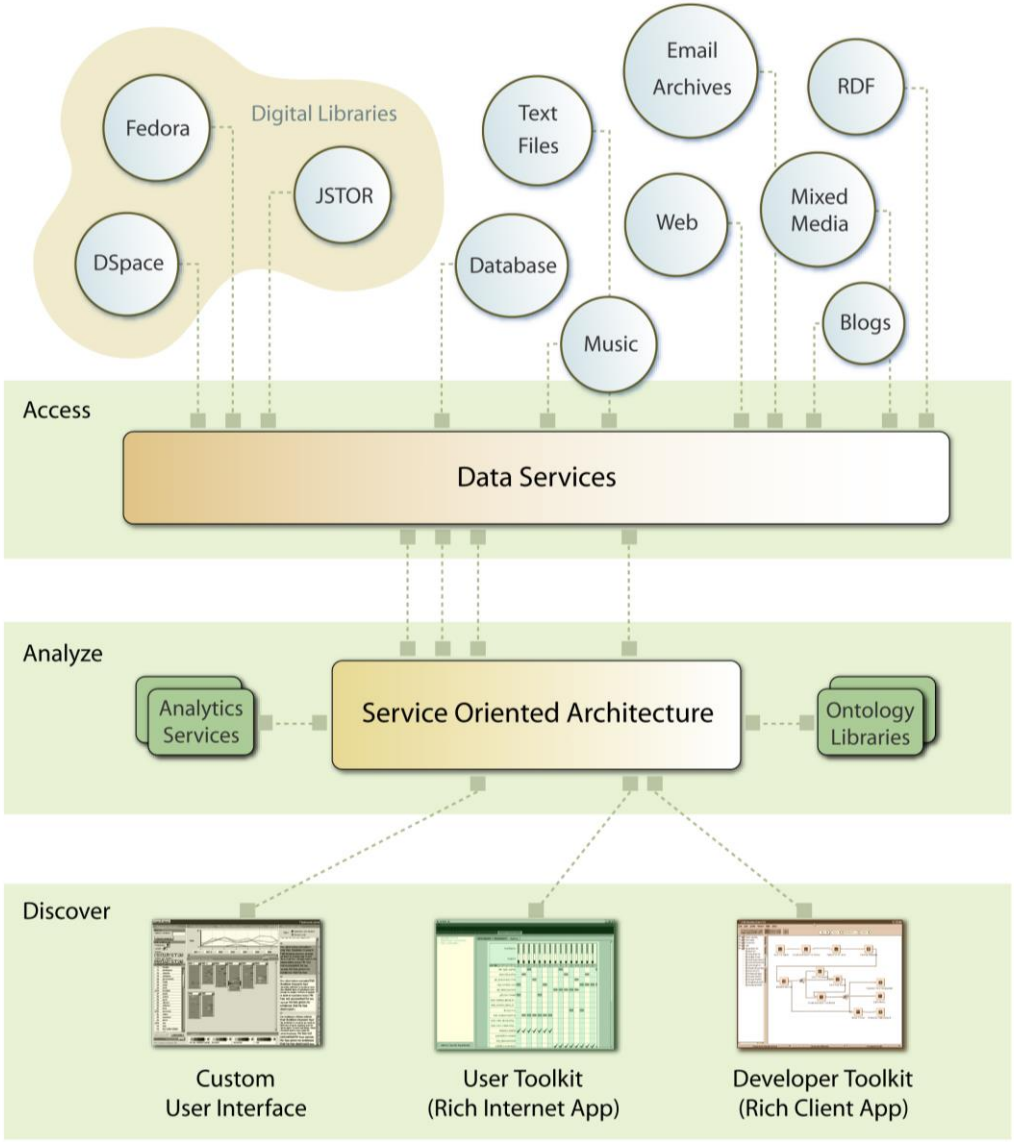
Framework for (re) using code

Works with variety of data formats

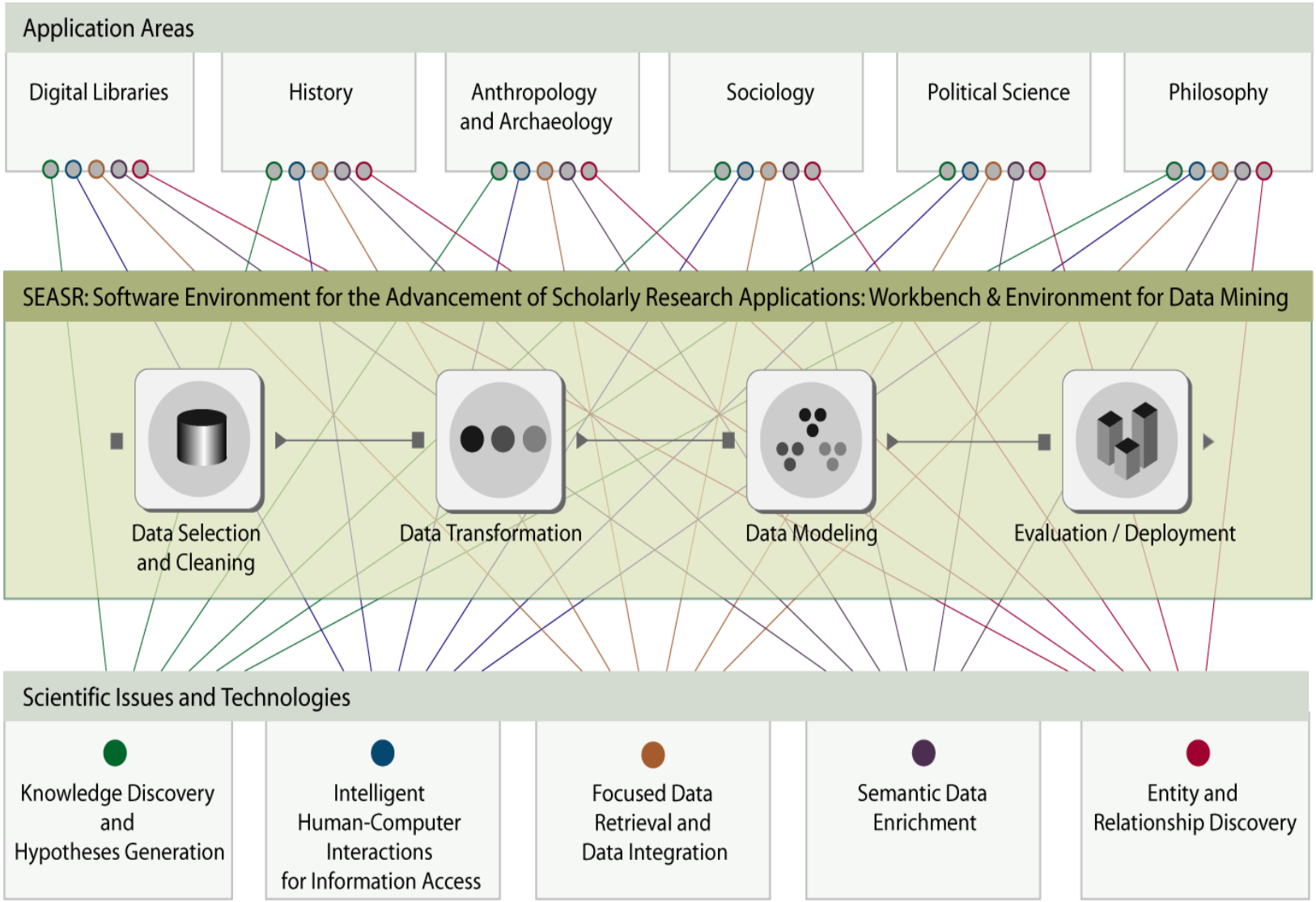
Information Analysis Components/Flows

Based on Semantic Concepts

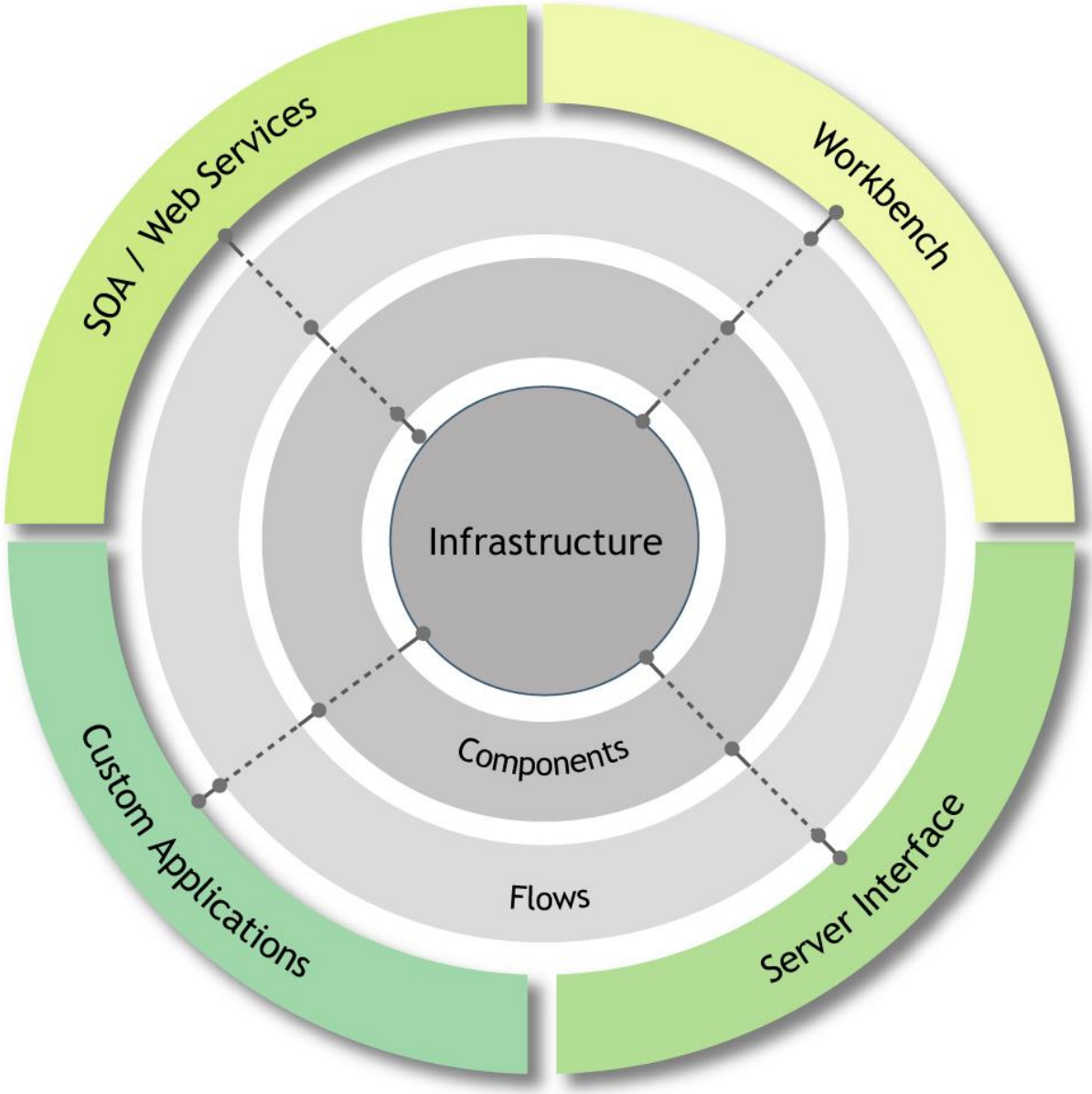
The SEASR Picture



SEASR Overview



SEASR Architecture



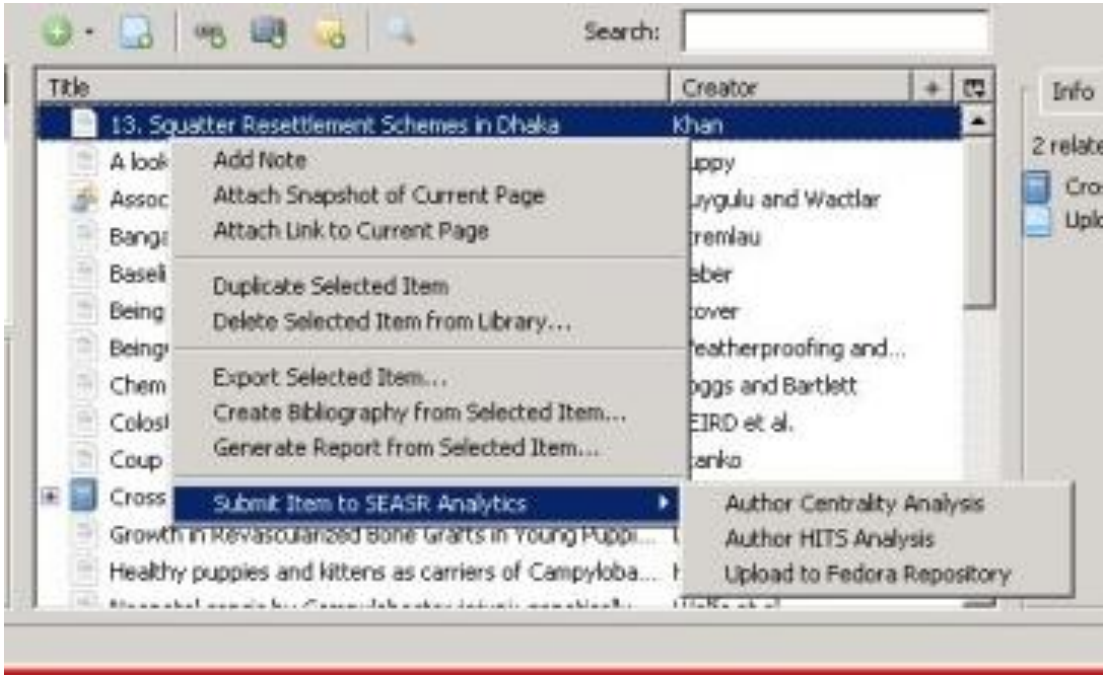
SEASR @ Work – Zotero

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



Community Hub

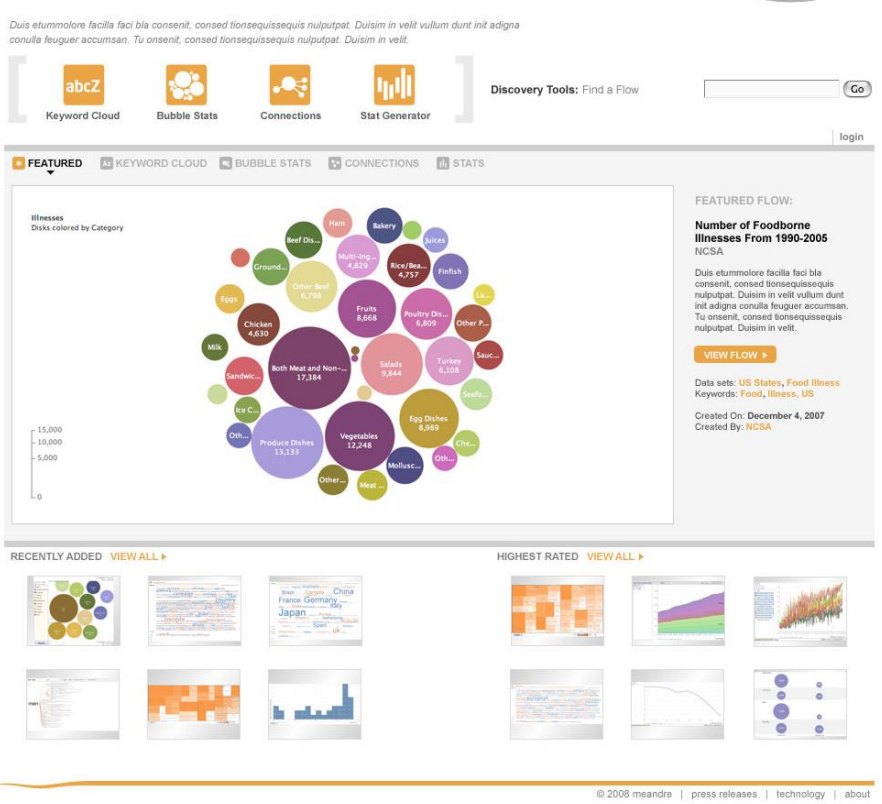
Explore existing flows to find others of interest

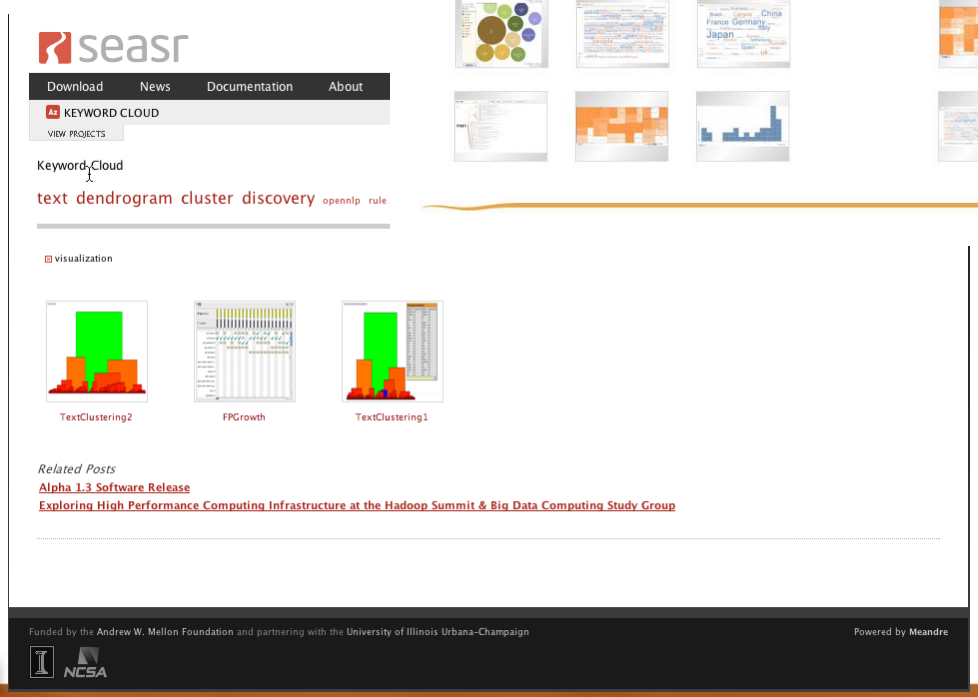
- Keyword Cloud
- Connections

Find related flows

Execute flow

Comments





seasr central

feedback | login | search

Categories Recently Added Top 50 Submit About RSS

Featured Component [read more]

Word Counter *by Jane Doe*

Description

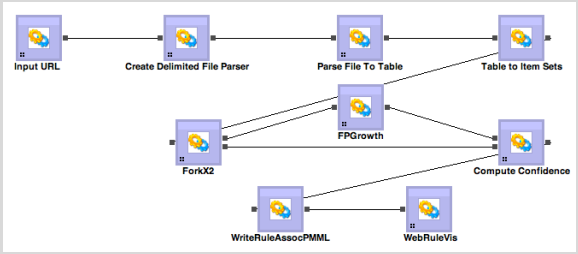
Amazing component that given text stream, counts all the different words that appear on the text

Rights:

NCSA/UofI open source license

Featured Flow [read more]

FPGrowth *by Joe Does*



Browse

Type	Categories	Name
Component	Image	Author Centrality
Flows	JSTOR	Readability
	Zotero	Upload Fedora

By Joe Doe
Rights: NCSA/UofI
Description: Webservices given a Zotero entry tries to retrieve the content and measure its

SEASR Central: Use Cases

- register for an account
- search for components / flows
- browse components / flows / categories
- upload component / flow
- share component / flow with: everyone or group
- unshare component / flow
- create group / delete group
- join group / leave group
- create collection
- generate location URL (permalink) for components, flows, collection (the location URL can be used inside the Workbench to gain access to that component or flows)
- view latest activity in public space / my groups

Community Hub: Connections Design

Funding Text-Mining in the Humanities

Andrew W. Mellon Foundation: Nora, WordHoard, MONK projects, 2004-2009

National Endowment for the Humanities:
Digging Into Data
http://www.diggingintodata.org/

Supercomputing in the humanities:
http://newscenter.lbl.gov/feature-stories/2008/12/22/humanitiesnersc/
https://www.sharcnet.ca/my/research/hhpc
http://www.ncsa.illinois.edu/News/09/0625NCSAICHASS.html

References

- *D-Lib Magazine*: http://www.dlib.org/
- "Sacred Electronics," *Time*: http://bit.ly/PPaAR
- *Digital Humanities Quarterly*: http://digitalhumanities.org/dhq/
- DH 2009: http://www.mith2.umd.edu/dh09/
- Bei Yu: http://www.beiyu.info/
- Philomine: http://philologic.uchicago.edu/philomine/
- The MONK Project: http://www.monkproject.org/
- SEASR: http://www.seasrproject.org/

unsworth@illinois.edu