The Center for Research Libraries

Text-Mining Opportunities and Challenges

November 13, 2012

CRL Global Resources Forum

Analysis

Intelligence

Library investment

Today’s agenda

- **Recent Trends in Text-Mining and Library Services: Research Library Perspective**

  David Magier, Princeton University Library

- **Elsevier Perspective on Mining the Scientific Literature**

  Judson Dunham, Elsevier

- **Gale’s Support of Text- and Data-Mining Requests**

  Ray Abruzzi, Cengage Learning

- **Text-Mining in the Cloud**

  Ann Okerson, Center for Research Libraries

Presenters

- **David Magier**

Associate University Librarian
for Collection Development

Princeton University Library

PRINCETON UNIVERSITY LIBRARY

Recent Trends in Text Mining and Library Services: Research Library Perspective

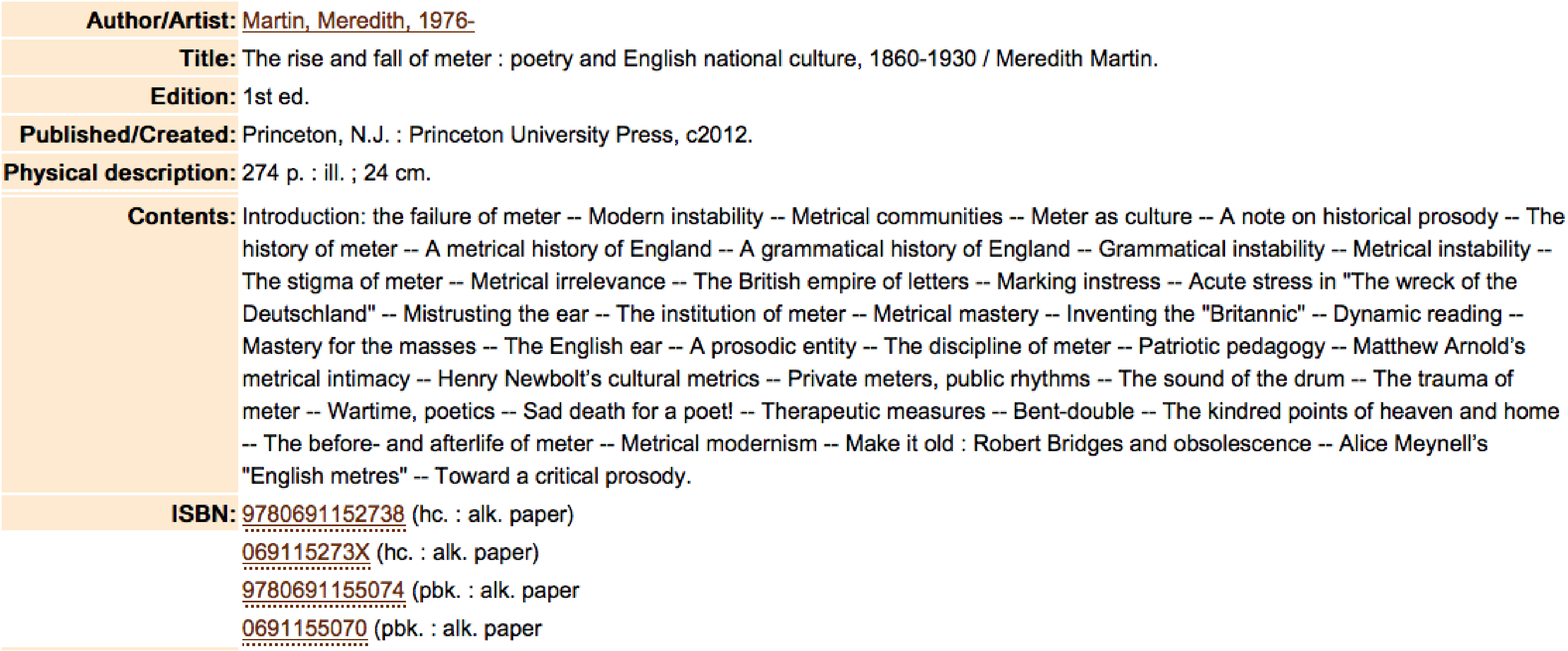
David Magier
Assoc. Univ. Librarian for Collection Development
Princeton University Library

What we do with text

- Use it to enable more granular **discovery**
  - to find and read specific chunks of content
  - "consumptive" use

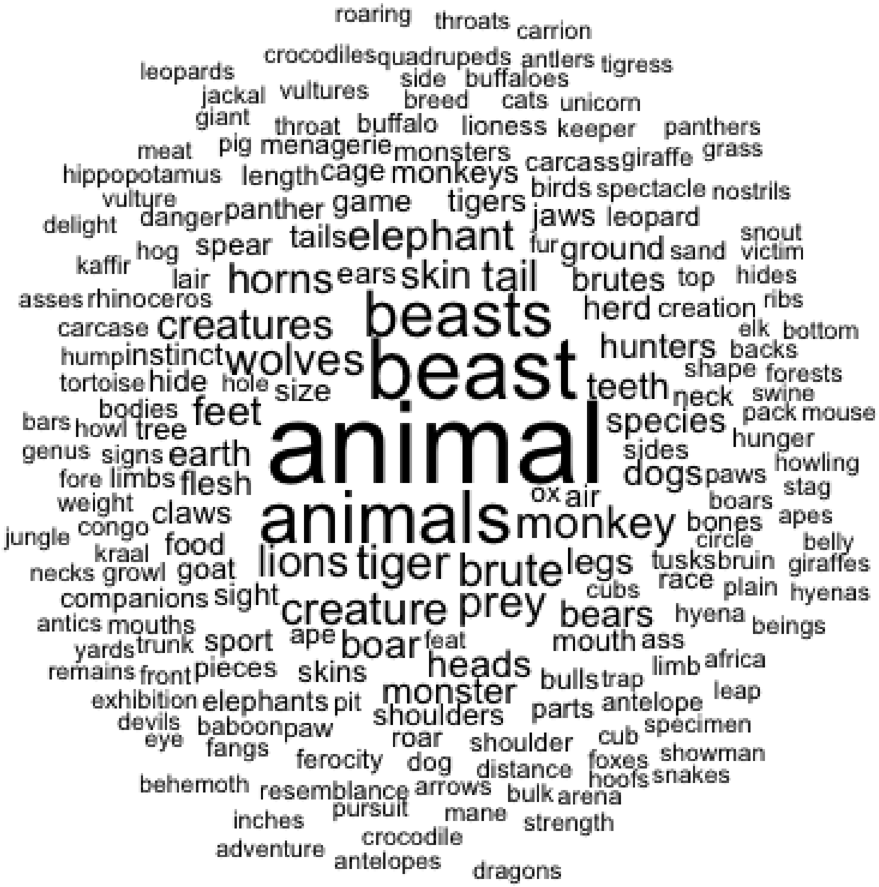
- Use lots of it to analyze large corpora of text
  - subject text to statistical analyses to discover **meaningful correlations and trends**
  - "non-consumptive" use

Examples



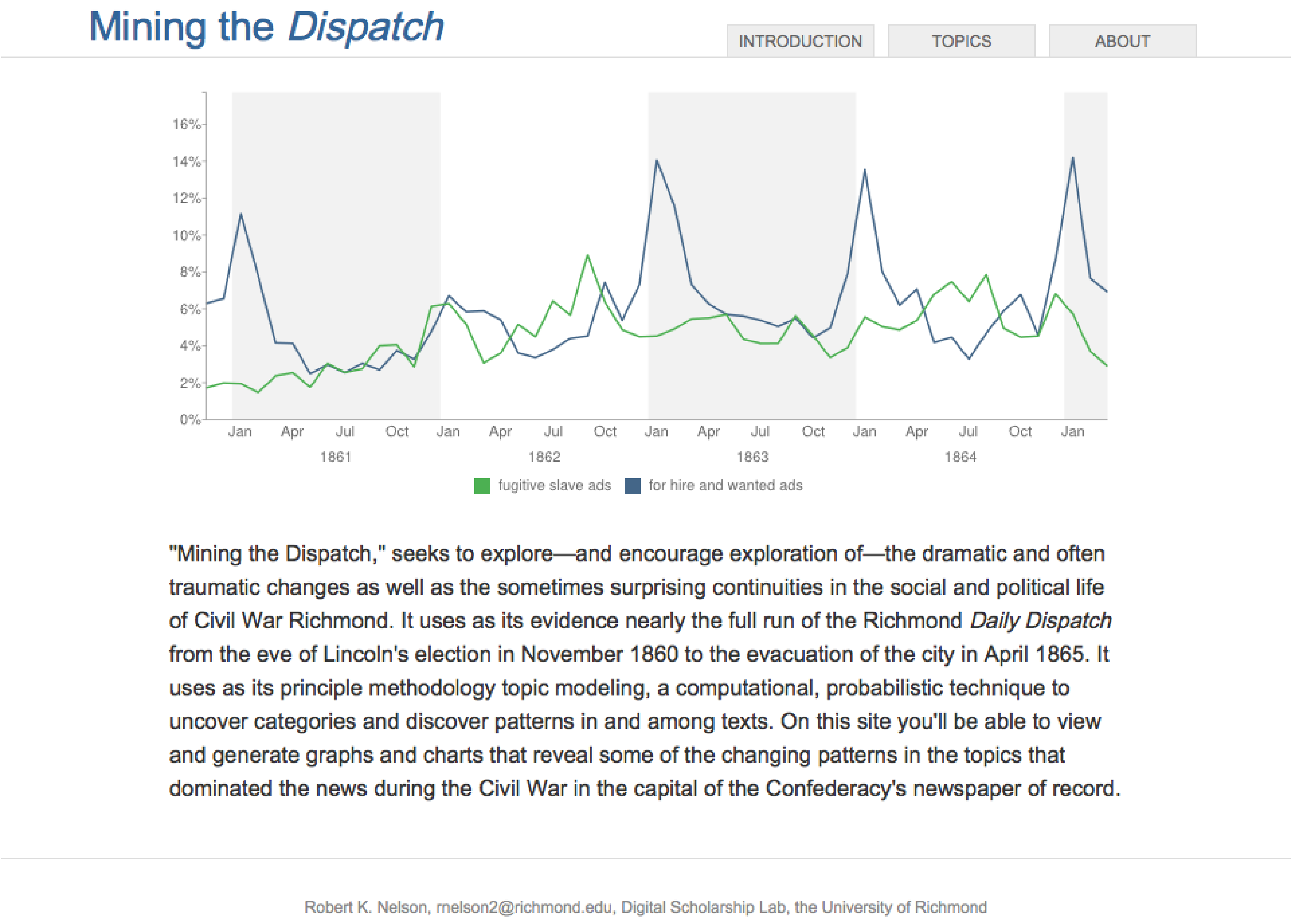
Author/Artist:	Martin, Meredith, 1976-
Title:	The rise and fall of meter : poetry and English national culture, 1860-1930 / Meredith Martin.
Edition:	1st ed.
Published/Created:	Princeton, N.J. : Princeton University Press, c2012.
Physical description:	274 p. : ill. ; 24 cm.
Contents:	Introduction: the failure of meter -- Modern instability -- Metrical communities -- Meter as culture -- A note on historical prosody -- The history of meter -- A metrical history of England -- A grammatical history of England -- Grammatical instability -- Metrical instability -- The stigma of meter -- Metrical irrelevance -- The British empire of letters -- Marking instress -- Acute stress in "The wreck of the Deutschland" -- Mistrusting the ear -- The institution of meter -- Metrical mastery -- Inventing the "Britannic" -- Dynamic reading -- Mastery for the masses -- The English ear -- A prosodic entity -- The discipline of meter -- Patriotic pedagogy -- Matthew Arnold's metrical intimacy -- Henry Newbolt's cultural metrics -- Private meters, public rhythms -- The sound of the drum -- The trauma of meter -- Wartime, poetics -- Sad death for a poet! -- Therapeutic measures -- Bent-double -- The kindred points of heaven and home -- The before- and afterlife of meter -- Metrical modernism -- Make it old : Robert Bridges and obsolescence -- Alice Meynell's "English metres" -- Toward a critical prosody.
ISBN:	9780691152738 (hc. : alk. paper) 069115273X (hc. : alk. paper) 9780691155074 (pbk. : alk. paper 0691155070 (pbk. : alk. paper

Matthew Jockers:
500 Themes from a corpus of 19th-Century Fiction



Macroanalysis: Digital Methods and Literary History (UIUC Press, 2013),
http://www.matthewjockers.net/macroanalysisbook/macro-themes/

Robert Nelson: Mining the Dispatch



Mining the *Dispatch*

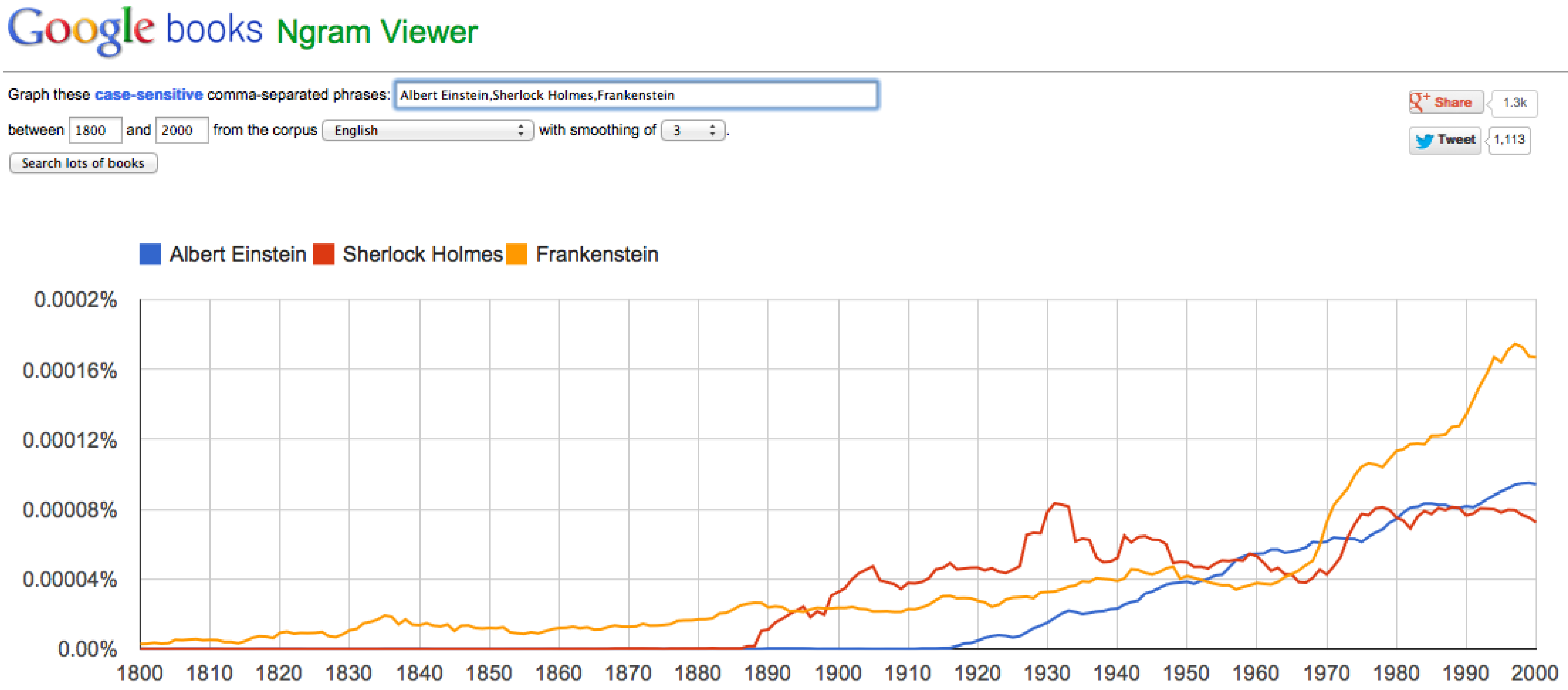
"Mining the Dispatch," seeks to explore—and encourage exploration of—the dramatic and often traumatic changes as well as the sometimes surprising continuities in the social and political life of Civil War Richmond. It uses as its evidence nearly the full run of the Richmond *Daily Dispatch* from the eve of Lincoln's election in November 1860 to the evacuation of the city in April 1865. It uses as its principle methodology topic modeling, a computational, probabilistic technique to uncover categories and discover patterns in and among texts. On this site you'll be able to view and generate graphs and charts that reveal some of the changing patterns in the topics that dominated the news during the Civil War in the capital of the Confederacy's newspaper of record.

Robert K. Nelson, rnelson2@richmond.edu, Digital Scholarship Lab, the University of Richmond

http://dsl.richmond.edu/dispatch/

The Corpus:

GoogleBooks Ngram Viewer



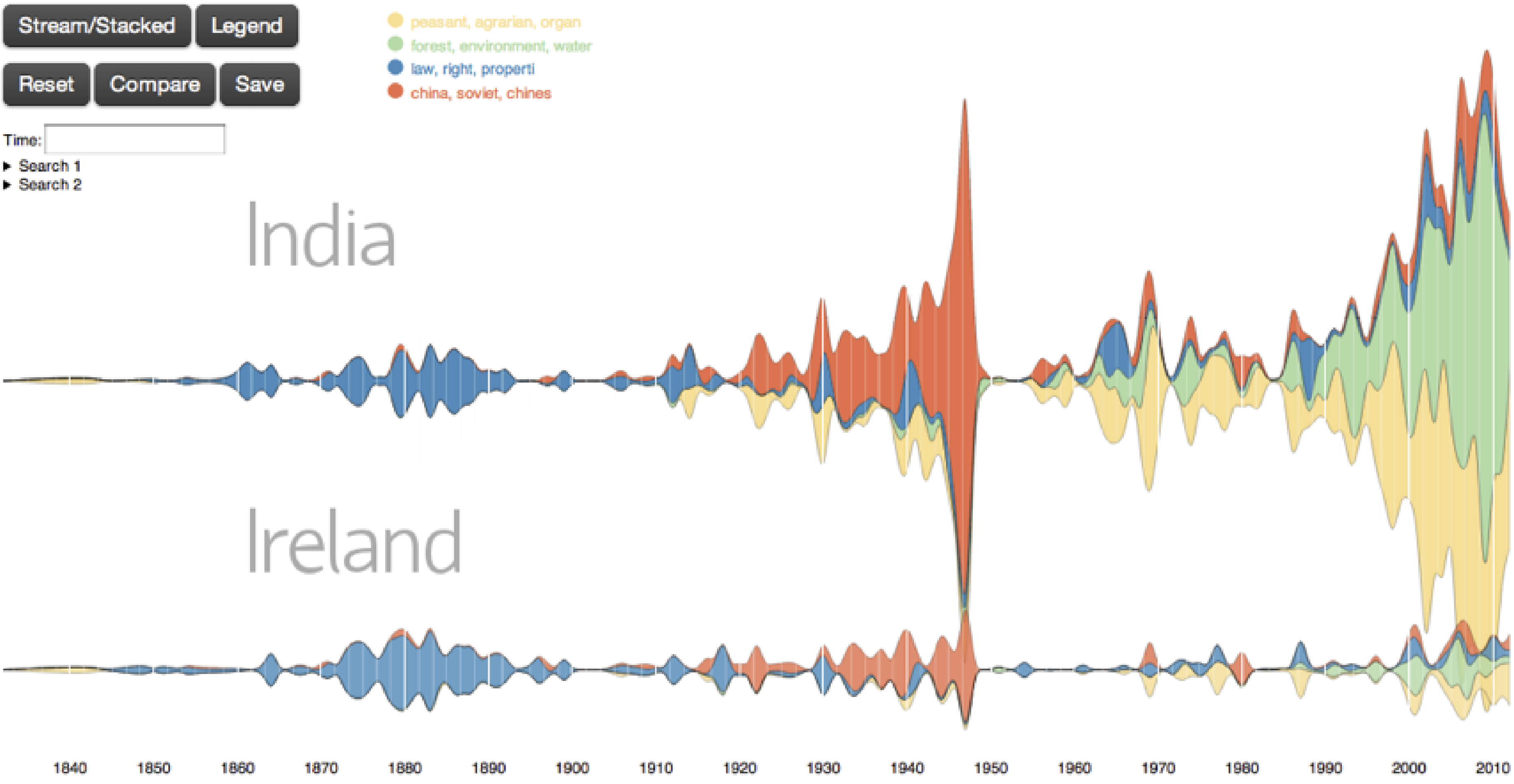
Google books Ngram Viewer

Graph these **case-sensitive** comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein

between 1800 and 2000 from the corpus English with smoothing of 3.

http://books.google.com/ngrams

Paper Machines:
a text-mining visualization plug-in for Zotero from Harvard's MetaLab

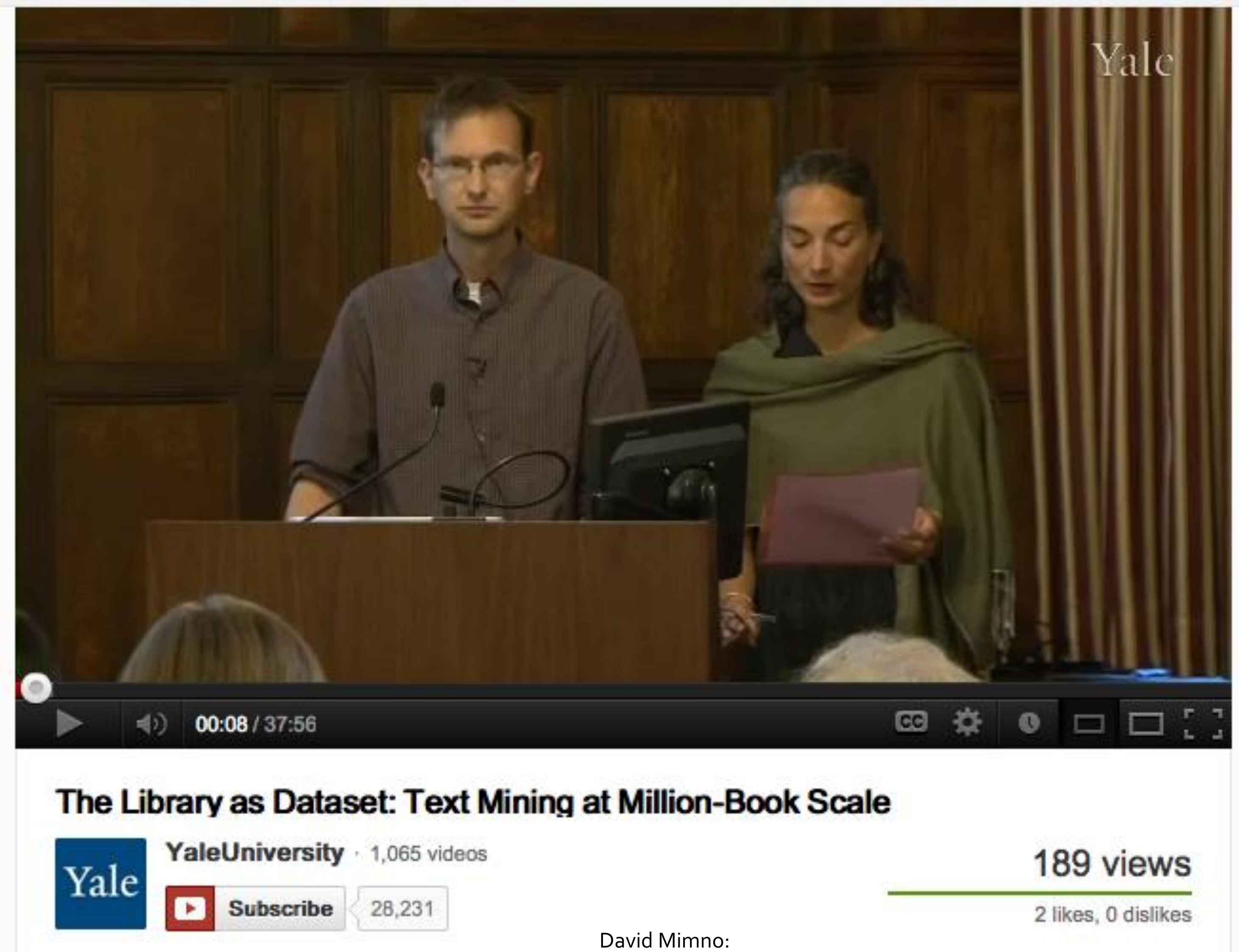


Four topics shown for all documents mentioning India (top) and/or Ireland (bottom).

http://metalab.harvard.edu/2012/07/paper-machines/

The Corpus:

The Library as Data-Set: Text Mining at Million-Book Scale



The Library as Dataset: Text Mining at Million-Book Scale

YaleUniversity · 1,065 videos

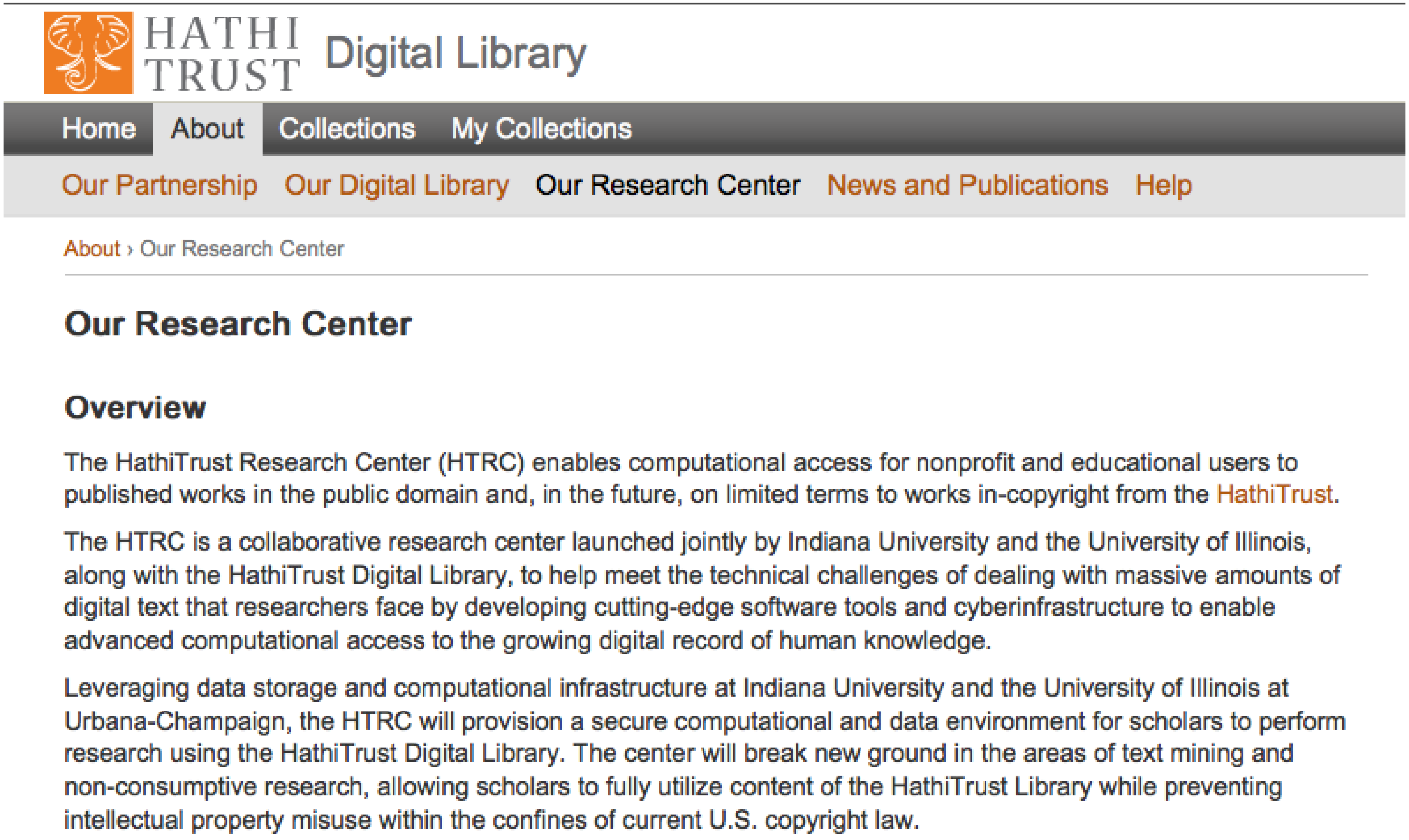
189 views

2 likes, 0 dislikes

David Mimno:
http://www.youtube.com/watch?v=o_j9jinLRIQ

The Corpus:

HathiTrust Research Center



Our Research Center

Overview

The HathiTrust Research Center (HTRC) enables computational access for nonprofit and educational users to published works in the public domain and, in the future, on limited terms to works in-copyright from the HathiTrust.

The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge.

Leveraging data storage and computational infrastructure at Indiana University and the University of Illinois at Urbana-Champaign, the HTRC will provision a secure computational and data environment for scholars to perform research using the HathiTrust Digital Library. The center will break new ground in the areas of text mining and non-consumptive research, allowing scholars to fully utilize content of the HathiTrust Library while preventing intellectual property misuse within the confines of current U.S. copyright law.

http://www.hathitrust.org/htrc

Where does the library fit?

- Who creates tools?
- How does content become “a corpus”?
- **Who can connect the scholars to the content AND the tools they need?**

Presenters

- **Judson Dunham**

  Senior Product Manager, Elsevier

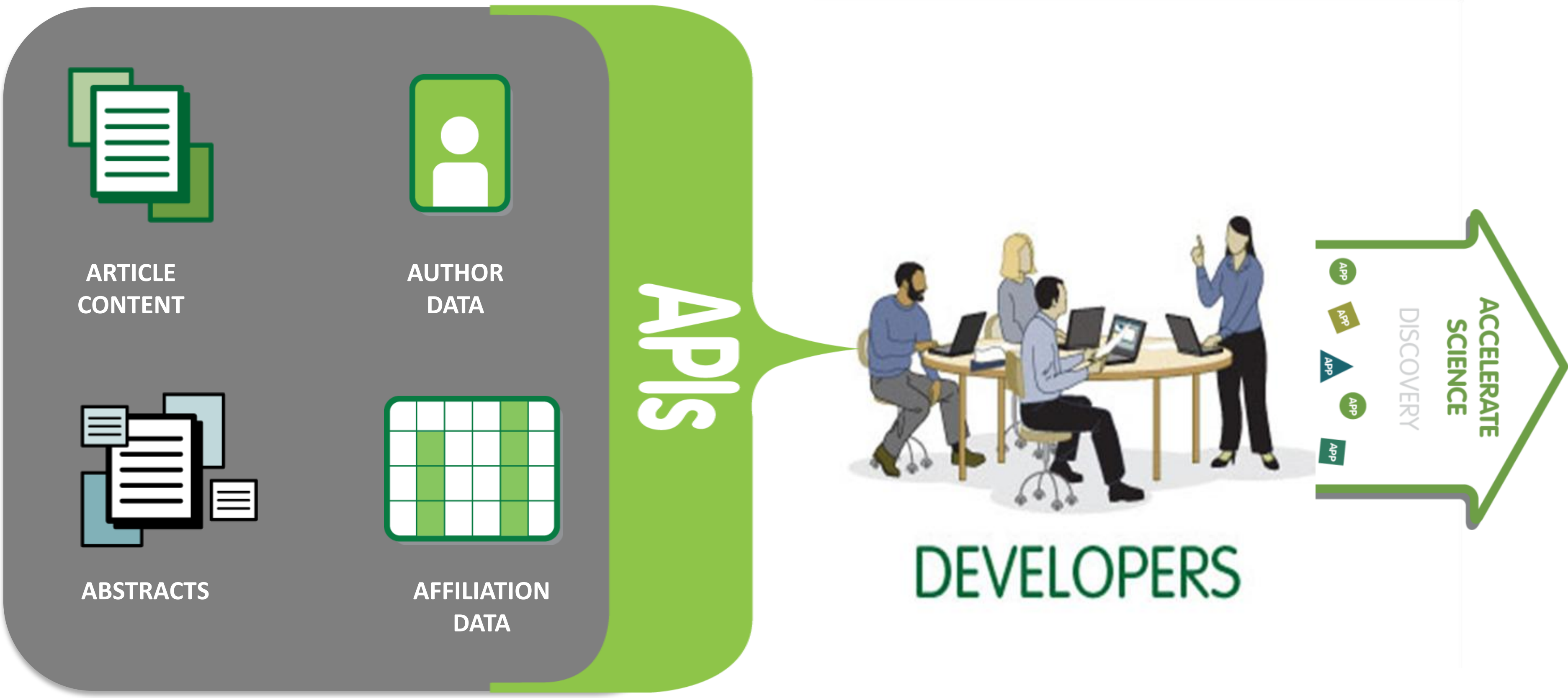
Elsevier Perspective on Mining the Scientific Literature

Judson Dunham
Sr. Product Manager, Elsevier

2010

http://vignettebricks.blogspot.com/2007/07/miss-librarian-in-library-with-slippery.html

Elsevier content and APIs for research community



ARTICLE CONTENT

AUTHOR DATA

ABSTRACTS

AFFILIATION DATA

APIs

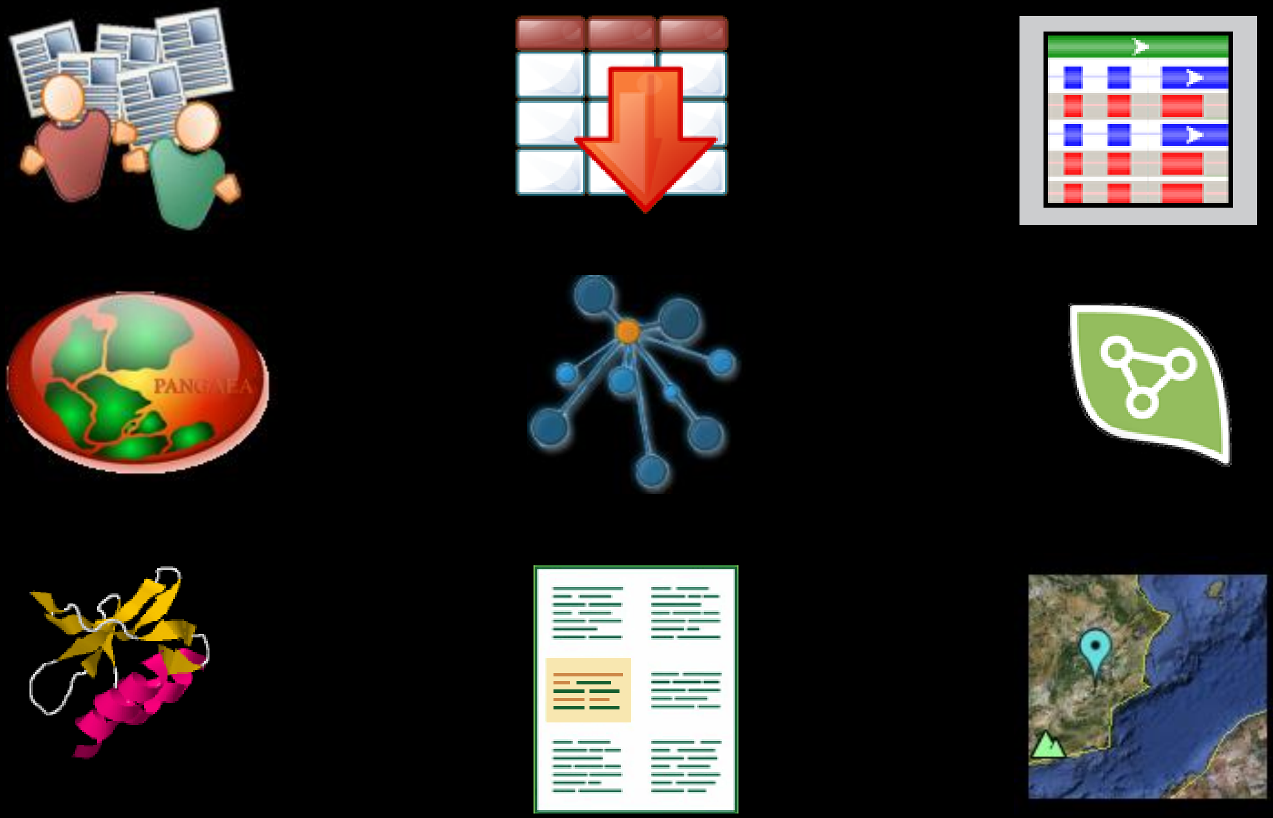
DEVELOPERS

ACCELERATE SCIENCE DISCOVERY

Collaboration at a different level

2011

RESEARCH WITH APPS



2012

What researchers want from content mining

- Assurance of right to mine subscribed or freely available content across publishers
- Clear understanding of what they can do with the outputs
- Standardized, mineable formatting of data
- Stable, reliable systems for retrieving and storing content

How Elsevier is working to address this

Simple process to support text mining

- Researchers/Librarians can request text mining access for specific projects on case by case basis
- This is a pilot period – there is still much to learn about researcher needs

Provide efficient methods for content retrieval

- **Access to APIs for individual researchers** via click through agreements on the web
- **Slimmed down content formats** optimized for text mining

How Elsevier is working to address this

Clear guidelines on reuse and redistribution

- Simple **CC-BY-NC** license on output when redistributing results of text mining in a research support or other non-commercial tool
- **DOI link back to mined article** whenever feasible when displaying extracted content
- Clear guidelines and permissions on **snippets around extracted entities** to allow context when presenting results

USE CASE: Studying patterns of data sharing and reuse in the published literature

Text mining small subset of content

- Using literature as a data source to explore a research issue (in this case, how is data in repositories shared, referenced and reused in the literature)
- Requires 1,000’s of documents (not 1,000,000’s)
- Will result in papers published in journals, as well as a publicly available database providing information about reuse of specific datasets (mined from the literature)

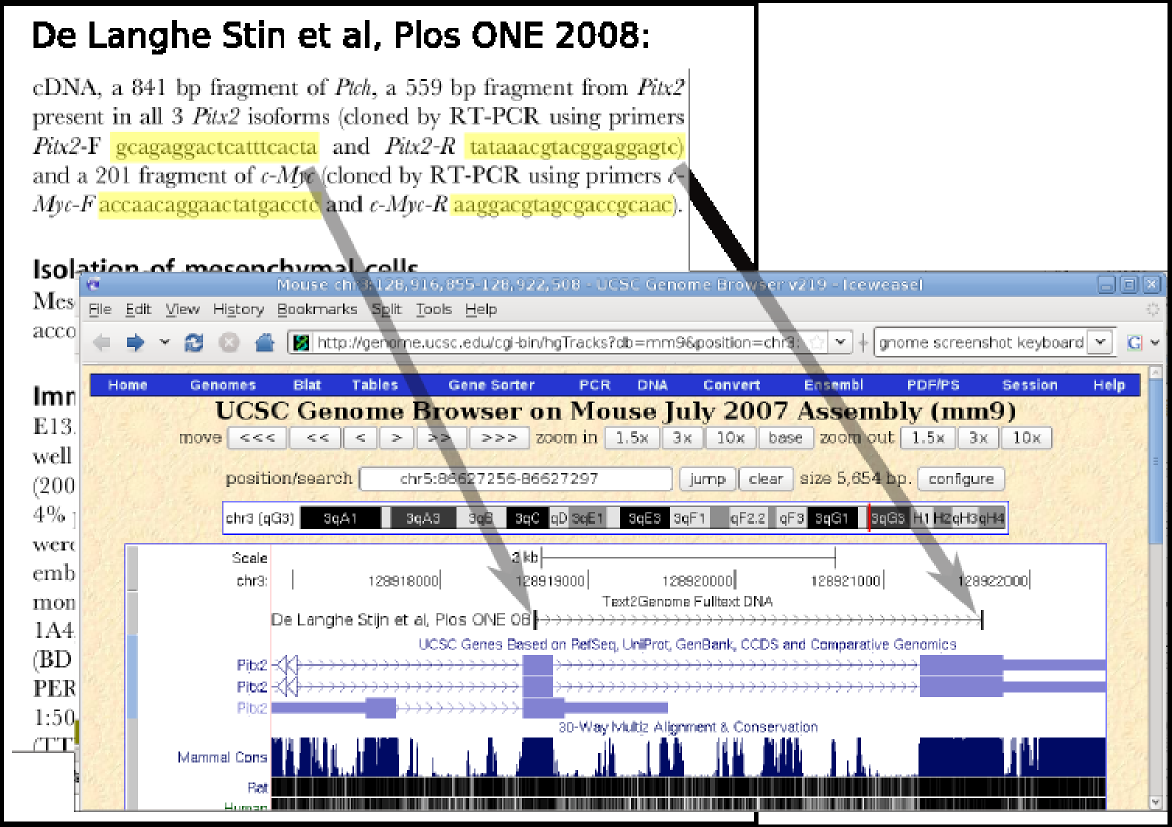
Solution: Access to subscribed content via API’s, permission to reuse results in non-commercial tool

Heather Piwowar - Postdoc at Duke University

USE CASE: Extraction of DNA sequences from millions of papers to facilitate biocuration and discovery

Text mining large volume of content

Identify all words in full-text scientific articles resembling DNA sequences, which are extracted and then mapped to public genome sequences. They can then be displayed on genome browser websites and used in data-mining applications.

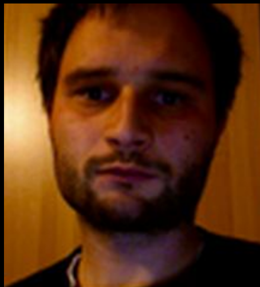


De Langhe Stin et al, Plos ONE 2008:

cDNA, a 841 bp fragment of *Ptch*, a 559 bp fragment from *Pitx2* present in all 3 *Pitx2* isoforms (cloned by RT-PCR using primers *Pitx2*-F gcagaggactcatttcacta and *Pitx2-R* tataaacgtacggaggagtc) and a 201 fragment of *c-Myc* (cloned by RT-PCR using primers *c-Myc-F* accaacaggaactatgacctc and *c-Myc-R* aaggacgtagcgaccgcaac).

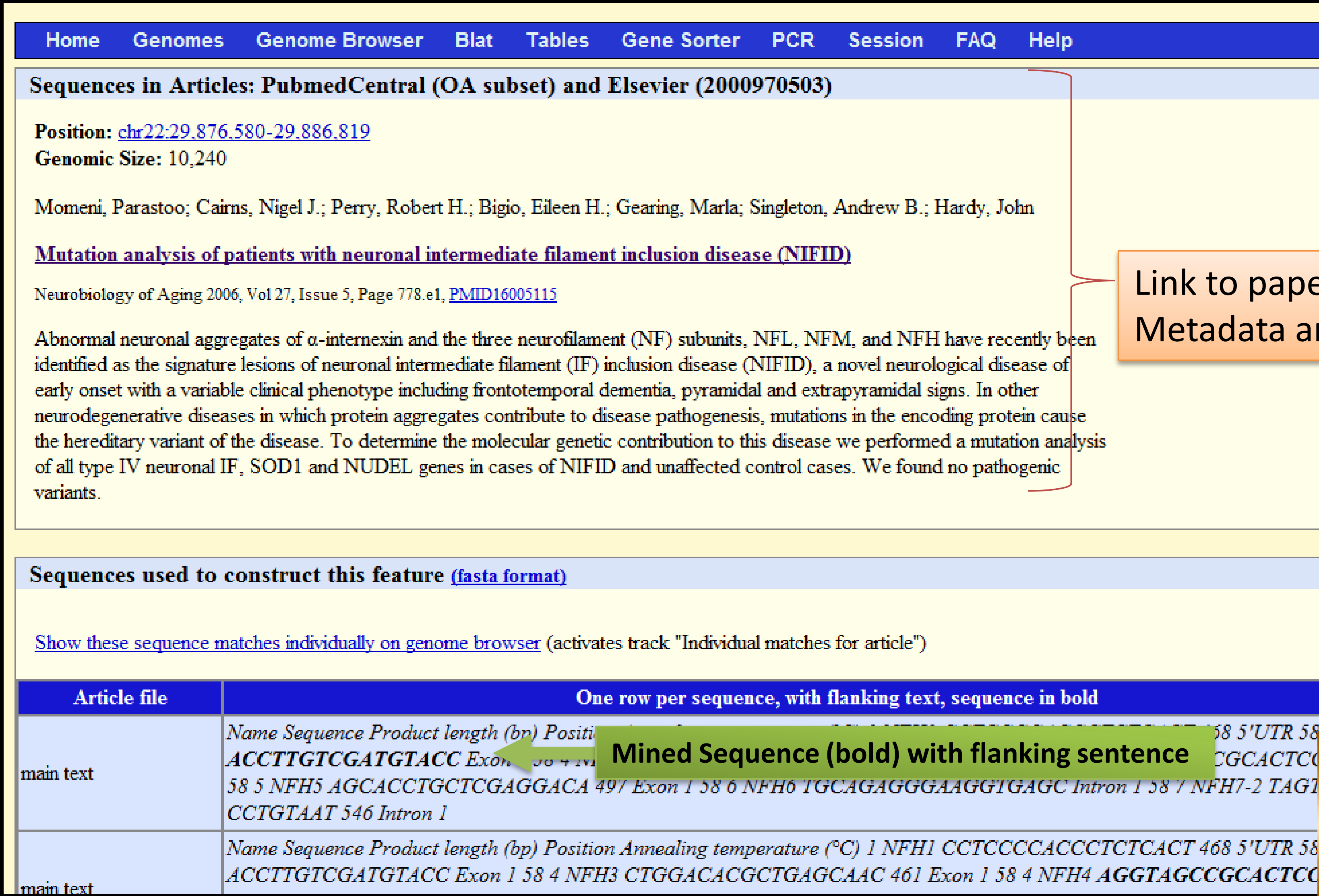
Isolation of mesenchymal cells

Mouse chr3:128,916,855-128,922,508 - UCSC Genome Browser v219 - Iceweasel
File Edit View History Bookmarks Split Tools Help
http://genome.ucsc.edu/cgi-bin/hgTracks?db=mm9&position=chr3
gnome screenshot keyboard
Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl PDF/PS Session Help
UCSC Genome Browser on Mouse July 2007 Assembly (mm9)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr5:86627256-86627297 jump clear size 5,654 bp. configure
chr3 (qG3) 3qA1 3qA3 3qB 3qC qD 3qE1 3qE3 3qF1 qF2.2 qF3 3qG1 3qG3 H1 H2 qH3 qH4
Scale 2 kb
chr3: 128918000 128919000 128920000 128921000 128922000
Text2Genome Fulltext DNA
De Langhe Stijn et al, Plos ONE 08
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Pitx2
Pitx2
Pitx2
30-Way Multiz Alignment & Conservation
Mammal Cons
Rat
Human



Max Haeussler – Post Doc, UCSC

USE CASE: Extraction of DNA sequences from millions of papers to facilitate biocuration and discovery



Sequences in Articles: PubmedCentral (OA subset) and Elsevier (2000970503)

Position: chr22:29,876,580-29,886,819
Genomic Size: 10,240

Momeni, Parastoo; Cairns, Nigel J.; Perry, Robert H.; Bigio, Eileen H.; Gearing, Marla; Singleton, Andrew B.; Hardy, John

Mutation analysis of patients with neuronal intermediate filament inclusion disease (NIFID)

Neurobiology of Aging 2006, Vol 27, Issue 5, Page 778.e1, PMID16005115

Abnormal neuronal aggregates of α-internexin and the three neurofilament (NF) subunits, NFL, NFM, and NFH have recently been identified as the signature lesions of neuronal intermediate filament (IF) inclusion disease (NIFID), a novel neurological disease of early onset with a variable clinical phenotype including frontotemporal dementia, pyramidal and extrapyramidal signs. In other neurodegenerative diseases in which protein aggregates contribute to disease pathogenesis, mutations in the encoding protein cause the hereditary variant of the disease. To determine the molecular genetic contribution to this disease we performed a mutation analysis of all type IV neuronal IF, SOD1 and NUDEL genes in cases of NIFID and unaffected control cases. We found no pathogenic variants.

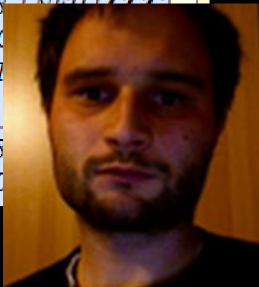
Link to paper with Metadata and abstract

Sequences used to construct this feature (fasta format)

Show these sequence matches individually on genome browser (activates track "Individual matches for article")

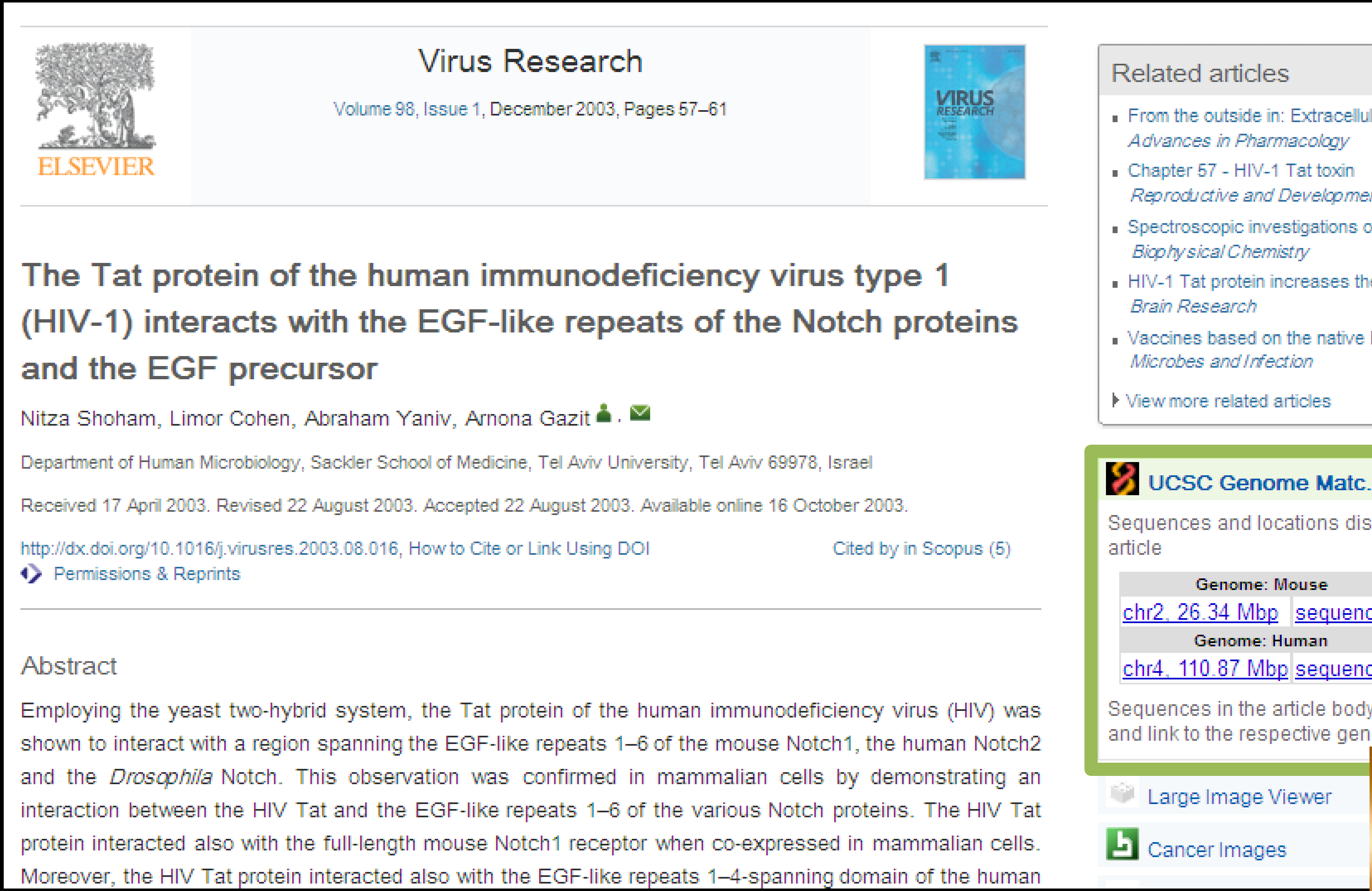
Article file	One row per sequence, with flanking text, sequence in bold
main text	*Name Sequence Product length (bp) Positi* … *68 5'UTR 58 2 NFH2 CCC* ***ACCTTGTCGATGTACC*** *Exon* … *CGCACTCC* *58 5 NFH5 AGCACCTGCTCGAGGACA 497 Exon 1 58 6 NFH6 TGCAGAGGGAAGGTGAGC Intron 1 58 7 NFH7-2 TAGT* *CCTGTAAT 546 Intron 1*
main text	*Name Sequence Product length (bp) Position Annealing temperature (°C) 1 NFH1 CCTCCCCACCCTCTCACT 468 5'UTR 58* *ACCTTGTCGATGTACC Exon 1 58 4 NFH3 CTGGACACGCTGAGCAAC 461 Exon 1 58 4 NFH4* ***AGGTAGCCGCACTCC***

Mined Sequence (bold) with flanking sentence



Max Haeussler – Post Doc, UCSC

USE CASE: Extraction of DNA sequences from millions of papers to facilitate biocuration and discovery



ELSEVIER

Virus Research

Volume 98, Issue 1, December 2003, Pages 57–61

The Tat protein of the human immunodeficiency virus type 1 (HIV-1) interacts with the EGF-like repeats of the Notch proteins and the EGF precursor

Nitza Shoham, Limor Cohen, Abraham Yaniv, Arnona Gazit

Department of Human Microbiology, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Received 17 April 2003. Revised 22 August 2003. Accepted 22 August 2003. Available online 16 October 2003.

http://dx.doi.org/10.1016/j.virusres.2003.08.016, How to Cite or Link Using DOI

Cited by in Scopus (5)

Permissions & Reprints

Abstract

Employing the yeast two-hybrid system, the Tat protein of the human immunodeficiency virus (HIV) was shown to interact with a region spanning the EGF-like repeats 1–6 of the mouse Notch1, the human Notch2 and the *Drosophila* Notch. This observation was confirmed in mammalian cells by demonstrating an interaction between the HIV Tat and the EGF-like repeats 1–6 of the various Notch proteins. The HIV Tat protein interacted also with the full-length mouse Notch1 receptor when co-expressed in mammalian cells. Moreover, the HIV Tat protein interacted also with the EGF-like repeats 1–4-spanning domain of the human

Related articles

- From the outside in: Extracellular activitie...
  Advances in Pharmacology
- Chapter 57 - HIV-1 Tat toxin
  Reproductive and Developmental Toxicology
- Spectroscopic investigations of HIV-1 trans-...
  Biophysical Chemistry
- HIV-1 Tat protein increases the permeability...
  Brain Research
- Vaccines based on the native HIV Tat protein...
  Microbes and Infection

View more related articles

UCSC Genome Matc...

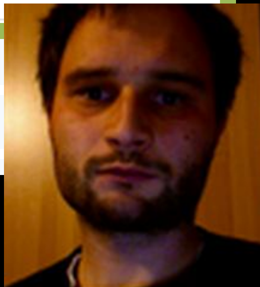
Sequences and locations discussed in this article

Genome: Mouse	
chr2, 26.34 Mbp	sequences
Genome: Human	
chr4, 110.87 Mbp	sequences

Sequences in the article body are highlighted and link to the respective genome.

Large Image Viewer

Cancer Images

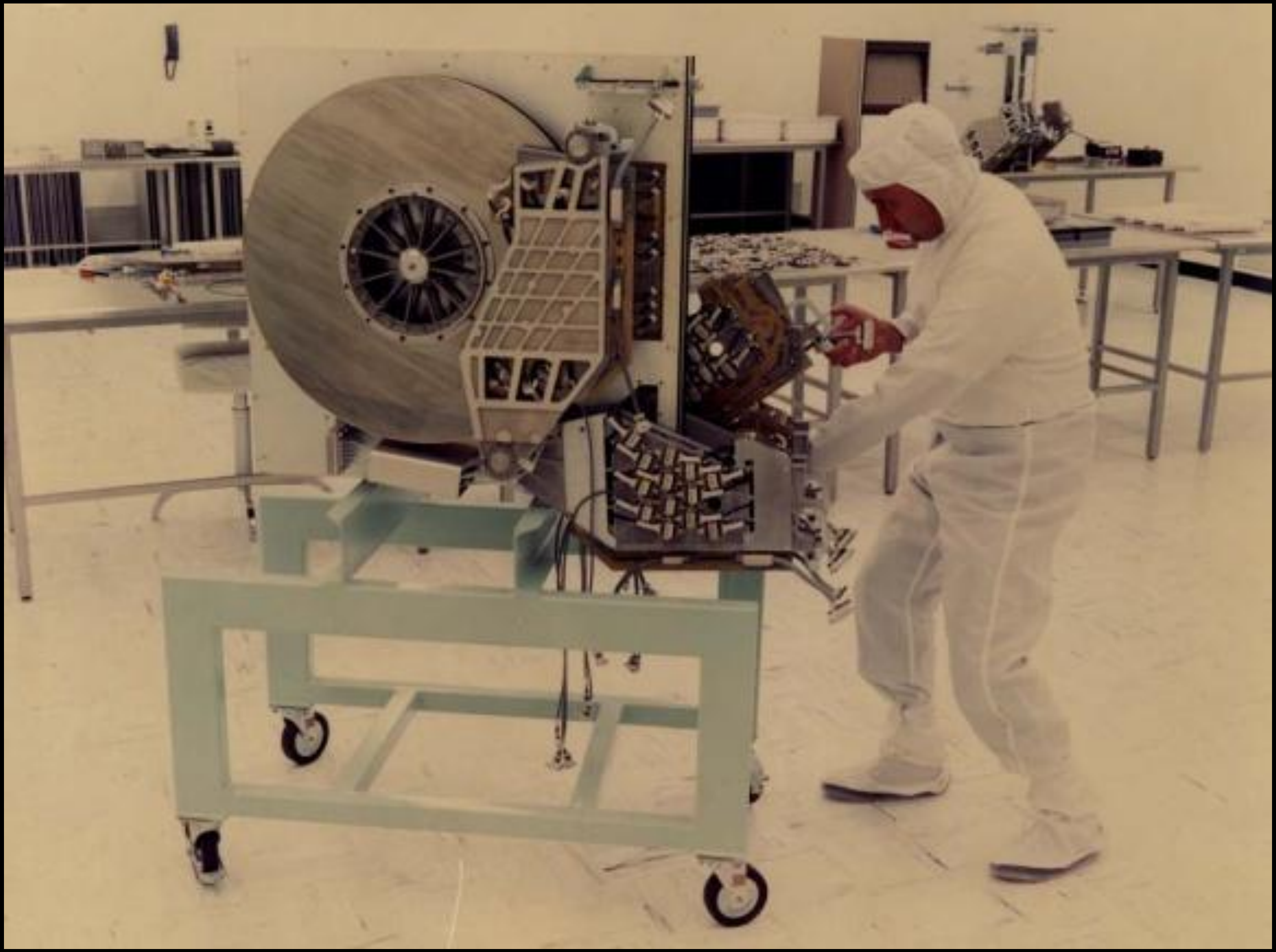


Max Haeussler – Post Doc, UCSC

Still some unsolved problems

Decentralized content aggregation:

- Retrieving and storing data locally = shipping loads of content around the web/world
- Still tons of formatting and preparation work for *every single text miner*
- Significant computational infrastructure required to do large scale text mining



A cloud-based service for running large-scale text mining without having to build infrastructure or worry about the "plumbing" would be valuable.

Still some unsolved problems

Preservation of knowledge:

- Storing and redistributing (valuable) text mining results requires long term infrastructure, hosting and other services
- Many academic research efforts end “when the postdoc moves on” – where does the mined knowledge go?
- **A service offering a reliable, persistent, managed, open venue for the hosting of content mining results would be valuable**

END

- j.dunham@elsevier.com/@judso

Thanks:

- Max Haeussler (UCSC)
- Heather Piwowar (UBC/Duke)
- Bradley Allen (Elsevier Labs)

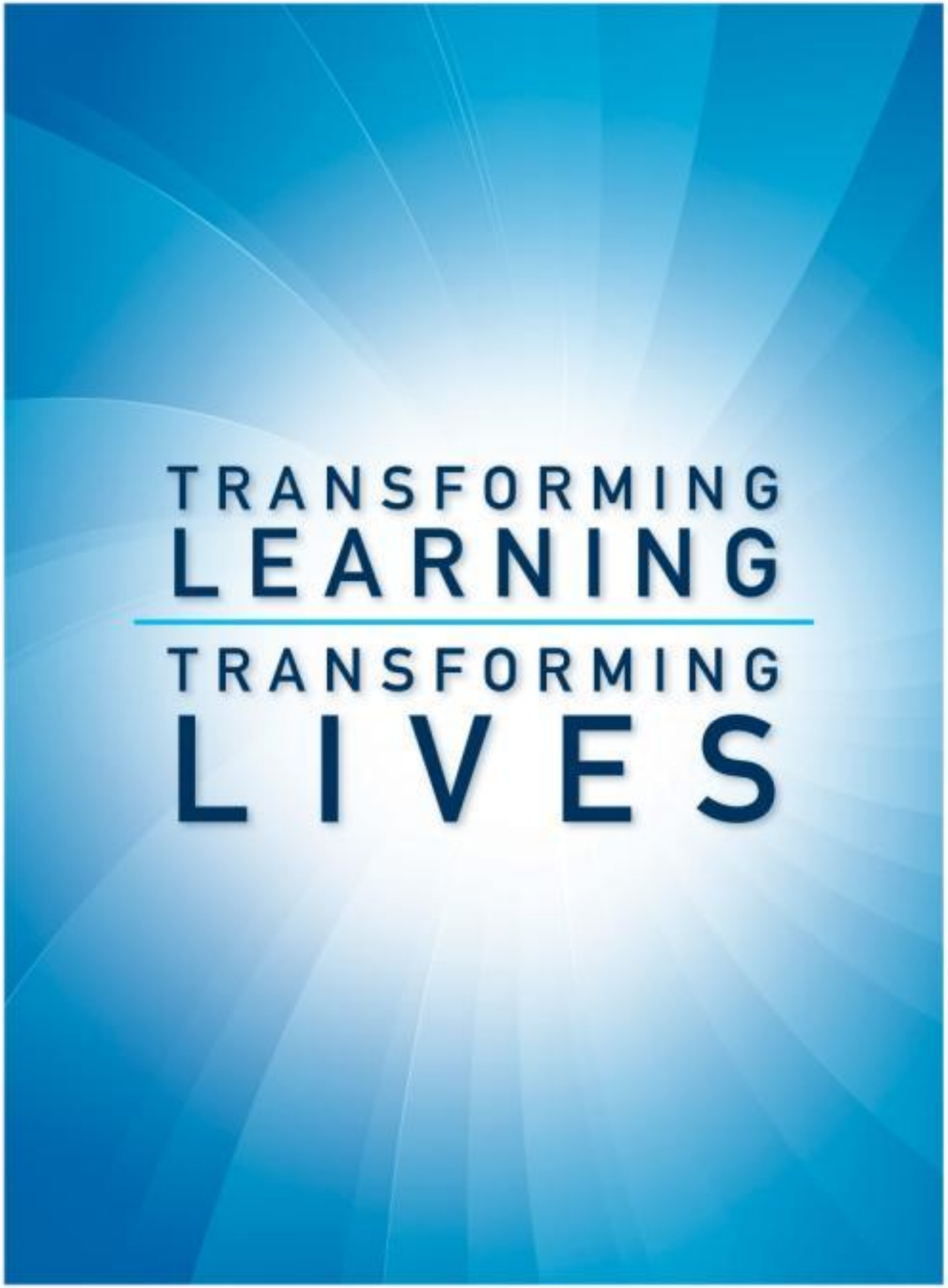
Presenters

- **Ray Abruzzi**

  Director, Strategic Planning

  Learning and Research Solutions

  Cengage Learning



CENGAGE Learning™

Ray Abruzzi
Director, Strategic Planning

Gale Scholarly Collaborations

- Gale has been involved with or is currently supporting over 30 projects to-date involving data-mining and/or textual analysis
- Efforts range from individuals to institutional to project-based groups
- Gale supports these efforts through a variety of means, most often providing:
  - Raw data consisting of entire collections, subsets of collections, or subsets of multiple collections, in multiple, specific formats
  - Technical advice and support
  - Connecting scholarly efforts through our involvement in multiple projects

Examples: IMPACT (Improving Access to Text)

Objectives: Significantly improve access to historical text

- Innovate OCR technology
  – By exploring the challenges using different approaches
  – By developing cutting-edge approaches such as collaborative correction
- Provide innovative language technologies to remove the historical language barrier
- Ensure the interoperability of the results
  – By defining an overall technical architecture and monitoring technical integration across all parts of the project
- Take away the barriers that stand in the way of the mass digitization of the European cultural heritage
- Provide Best Practice guidance about the operational context for digitization
- **Deliver a coherent program of dissemination, training and demonstration aimed at capacity-building in and beyond participating institutions**

Examples: IMPACT (Improving Access to Text)

Gale’s participation: Working with the British Library, IMPACT was provided with sample images and OCR’d text from the *19th Century British Library Newspapers*.

What they accomplished: This content formed part of the overall project dataset which was used by researchers to benchmark, develop/optimize and test language and OCR tools for historical material.

What the benefit was: This research led to increased OCR success for nine European languages/historical material and generated new areas of research interest.

Examples: 18th Connect (and NINES)

Objectives: NINES and 18thConnect are communities of scholars in 19th-century and 18th-century studies, respectively, organized around web sites that aggregate metadata and provide links to digital scholarly materials in their fields.

- In addition, for 18th-century studies, Gale has contributed page images (which are shown in snippets) and the OCR-generated texts from the ECCO collection to 18thConnect.
- These image snippets and the corresponding OCR text are made visible in 18thConnect's TypeWrite tool which allows 18thConnect's users to correct the OCR.
- Gale has given 18thConnect the opportunity to offer as a reward the full, typed texts (without page images) to anyone who corrects it. **Corrected texts may also be loaded back into ECCO to improve searching.**
- **18thConnect has also run new OCR engines on the page images from ECCO with the intent on creating improved OCR XML which could then be loaded back into the ECCO database, again resulting in a better search experience.**

Examples: 18th Connect and NINES

Gale’s participation: Provided OCR text and metadata for *Eighteenth Century Collections Online* and a range of 19th century content (some 19C material is metadata only).

What they accomplished: The metadata for ECCO is fully loaded and allows all scholars to search through it in conjunction with metadata from other scholarly databases. **The TypeWrite program is fully functional though no corrected texts have yet been returned to Gale.**

What the benefit was: At the current status of the projects, the 18thConnect and NINES websites provide a federated search-like approach to research across a range of metadata from key scholarly databases, allowing users to do one search instead of doing the same search in various standalone databases.

- If a user of these websites finds a work of interest within ECCO and their institution is a subscriber/owner of ECCO, they can then pass directly into the native application and see the entire work within ECCO.

A typical request

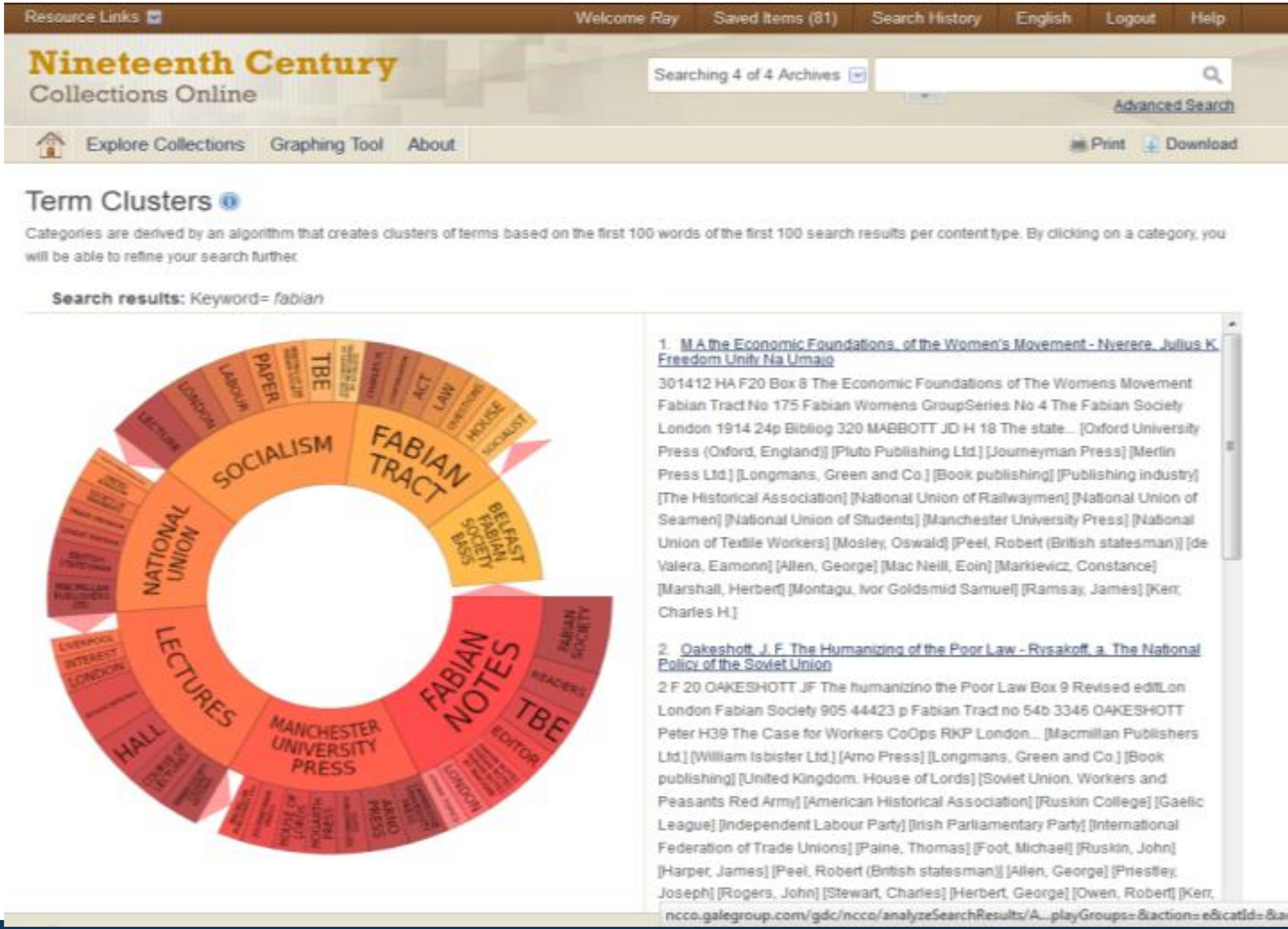
- “My name is XXX and I have a gig at XXX working for their Cultural Evolution of Religion project. I'm also the customer who is requesting access to the text files in the metadata of ECCO documents. See attached three images. Two are screenshots of my Advanced Search window, one of which I have highlighted each of the fields I have toggled. The only search term is 'earthquake', as you'll see. The difference between the two is that one isolates only documents in Religion and Philosophy whereas the other aims to capture all documents (in the same set of years) that contain 'earthquake'. The third image is a screenshot of my Gale Learning/ECCO user account page, which represents the search results in the these two searches. The narrow search issued 498 docs whereas the broad search issued 1453. Thanks for your help with this. If you can download the XML files from what I've given you now, I'll be very happy as it takes some time to click through on each page and copy and paste the Gale Doc ID numbers**. I need this by the end of the week to meet my project deadlines. Thank you.”**

Benefits to the academic community and to Gale

- **Gale is pleased to support academic efforts and projects of this nature**
  - **Of course, Gale is keen to generate goodwill in the academic community and to further scholarship**
- **The *expected outcomes* and *anticipated results* of these projects expand Gale's knowledge of both the changing needs of researchers (our customers) and broaden our understanding in terms of the kinds of tools Gale should be building into our resources**
  - **NCCO incorporates several new tools which enable a degree of textual analysis, and Gale is expanding these tools to our entire Digital Collections program over the next two years, starting with ECCO in 2013**
- **Gale may be able to improve or modify existing collections with improved data or tools based on the outcomes of various projects**

Leveraging these collaborations

- **Textual Analysis tools enable researchers to discover connections between documents, events, movements, and people.**



Nineteenth Century
Collections Online

Term Clusters

Categories are derived by an algorithm that creates clusters of terms based on the first 100 words of the first 100 search results per content type. By clicking on a category, you will be able to refine your search further.

Search results: Keyword= *fabian*

1. M A the Economic Foundations. of the Women's Movement - Nyerere, Julius K Freedom Unity Na Umajo

301412 HA F20 Box 8 The Economic Foundations of The Womens Movement Fabian Tract No 175 Fabian Womens GroupSeries No 4 The Fabian Society London 1914 24p Bibliog 320 MABBOTT JD H 18 The state... [Oxford University Press (Oxford, England)] [Pluto Publishing Ltd.] [Journeyman Press] [Merlin Press Ltd.] [Longmans, Green and Co.] [Book publishing] [Publishing industry] [The Historical Association] [National Union of Railwaymen] [National Union of Seamen] [National Union of Students] [Manchester University Press] [National Union of Textile Workers] [Mosley, Oswald] [Peel, Robert (British statesman)] [de Valera, Eamonn] [Allen, George] [Mac Neill, Eoin] [Markievicz, Constance] [Marshall, Herbert] [Montagu, Ivor Goldsmid Samuel] [Ramsay, James] [Kerr, Charles H.]

2. Oakeshott, J. F. The Humanizing of the Poor Law - Rysakoff, a. The National Policy of the Soviet Union

2 F 20 OAKESHOTT JF The humanizino the Poor Law Box 9 Revised editLon London Fabian Society 905 44423 p Fabian Tract no 54b 3346 OAKESHOTT Peter H39 The Case for Workers CoOps RKP London... [Macmillan Publishers Ltd.] [William Isbister Ltd.] [Arno Press] [Longmans, Green and Co.] [Book publishing] [United Kingdom. House of Lords] [Soviet Union. Workers and Peasants Red Army] [American Historical Association] [Ruskin College] [Gaelic League] [Independent Labour Party] [Irish Parliamentary Party] [International Federation of Trade Unions] [Paine, Thomas] [Foot, Michael] [Ruskin, John] [Harper, James] [Peel, Robert (British statesman)] [Allen, George] [Priestley, Joseph] [Rogers, John] [Stewart, Charles] [Herbert, George] [Owen, Robert] [Kerr,

ncco.galegroup.com/gdc/ncco/analyzeSearchResults/A...playGroups=&action=e&catId=&a

Challenges—Rights and Licenses

- **Rights: Gale licenses content from over 250 institutions to create our Digital Collections program**
  - **In many cases, this license permits very specific/limited uses of content and the data associated with that content**
  - **Gale often needs to seek permission to use this data in ANY project that does not fall under the provisions of the contract—institutions are reluctant to sign up for "any and all" uses for their content**
  - **Seeking permission or an amendment to the contract can take days, weeks, or months**
  - **Gale consistently seeks the broadest rights for our content licenses, with varying degrees of success**
  - **The process to acquire these rights for non-commercial projects-- which may or may not yield tangible, actionable benefits to Gale-- drains time and resources from other efforts**
  - **A license or contract with the person, institution or group is also needed, to ensure that the terms of use are not violated**

Challenges: Actionable outcomes, tangible benefits

- These are projects, often based on theories and experimentation—results may or may not yield tangible benefits
- Marginal improvements to searchability or OCR processes are not cost-justified. In other words, Gale cannot re-process millions of pages of data to make them "slightly" better
- Even when successful, results are often specific to the outcomes being sought, and are not transferable to any product of program offered from Gale
- The variety of data formats and tools and the processes utilized in these projects may not synch up with Gale—there may not be a ready means of ingestion for the outcomes, and the cost of creating one may far exceed the benefits of doing so

Summing up

- Gale continues to work with the digital humanities communities and support projects as we are able.
- The number and complexity of requests for data are growing, and Gale cannot donate resources to meet all of these requests
- Gale is actively seeking rights to use/utilize content (data) more broadly from source institutions to expedite our support of these projects
- Gale has built some initial tools for textual analysis and data-mining into our interfaces, and continues to expand on those tools to meet the needs of users at the point of research

Thank you—Questions?

CENGAGE Learning™

Ray Abruzzi, Director, Strategic Planning.
Ray.abruzzi@cengage.com

Presenters

Ann Okerson

Senior Advisor on Electronic Strategies,

Center for Research Libraries

Text Mining in the Cloud

Ann Okerson

Center for Research Libraries

Webinar - November 2012

property
quickly
access
every
lot
scholarly
thousands
before
tens
modern
big
sense
new
look
said
Intellectual
Trust
people
knows
scale
one
because
more
suggest
Dr
just
published
between
open
available
want
tool
journals
use
all
publishers
links
create
other
drugs
bought
Any
technique
publicly
find
gene
world
report
university
allow
allowed
ways
scientific
bn
through
genetics
science
scientists
Murray-Rust
DNA
sequence
last
similar
Elsevier
papers
already
Kiley
Professor
Bergman
Piwowar
able
Text2genome
diseases
never
year
help
individual
computers
powerful
genes
funded
know
sequences
led
things
patterns
Though
healthcare
articles
data
mining
literature
mining
information
potential
once
make
content
text
research
researchers

What we’ve been talking about

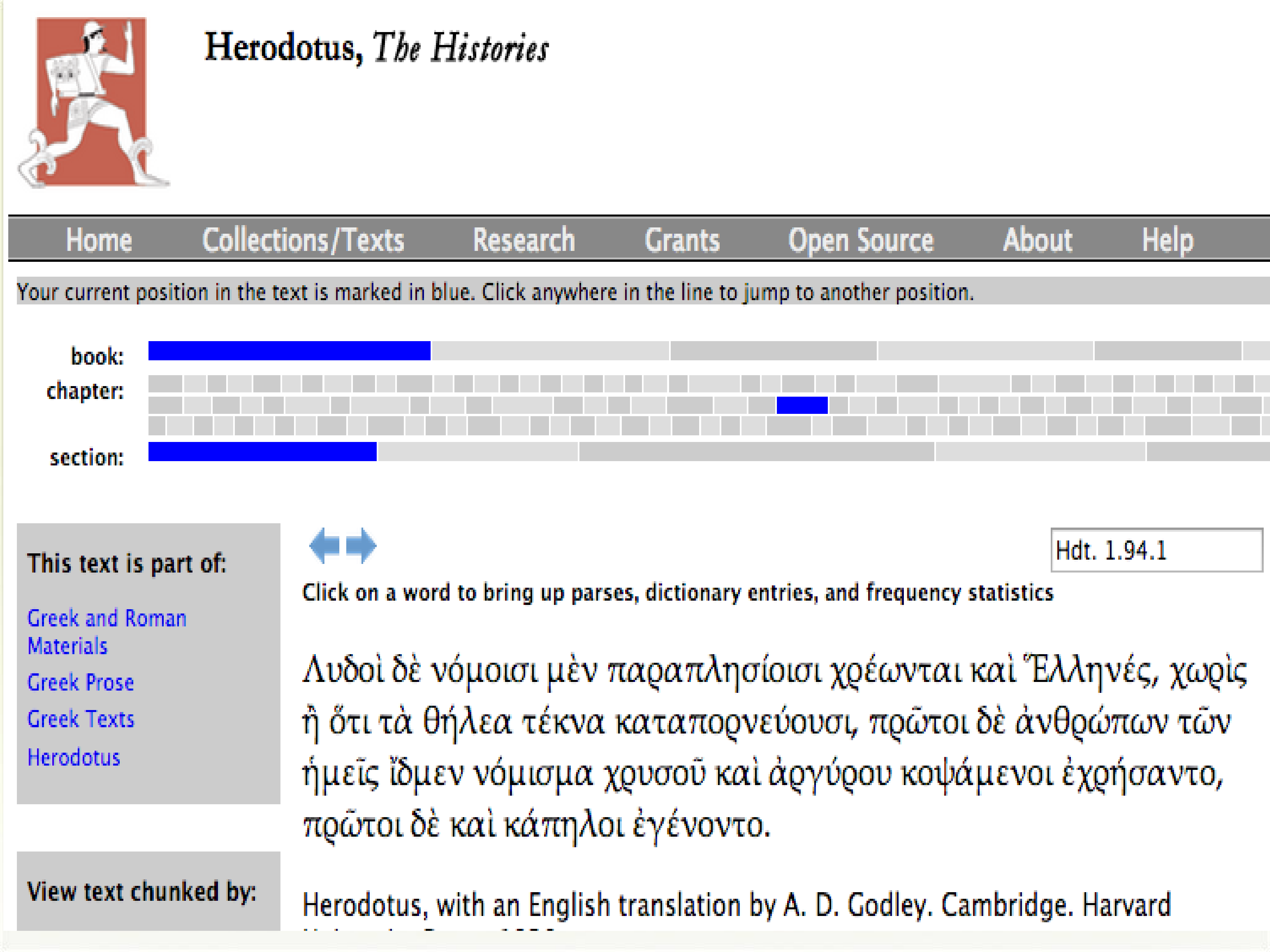
- “Automated processing of large amounts of digital data/content for purposes of information retrieval, extraction, interpretation, analysis” (Bernie Reilly)
- “Automated tools, techniques or technology to process large volumes of digital content that is often not well structured – to identify and select relevant information; to extract information from the content, to identify relationships within/between/across documents and incidents or events for meta-analysis” (Efke Smit)
- Access to data at a degree of granularity greater than an individual reader can make; and ability to process that granular data
- Mining terminology is a little vexing – text mining, data mining, content mining

Nothing new under the sun

- Don Swanson (1924-), American information scientist, known for his work in literature-based discovery in the biomedical domain
  - For some years, Dean of the U of Chicago Library School
- His method is called “Swanson linking:” connecting two pieces of knowledge previously thought to be unrelated
- Part of “Arrowsmith System project,” which sought to determine meaningful links between Medline articles
  - It may be known that illness A is caused by chemical B, and that drug C is known to reduce the amount of chemical B in the body
  - However, because the respective articles were published separately from one another (called "disjoint data"), the relationship between illness A and drug C may be unknown
- Don called this “undiscovered public knowledge”

Then, the next generation

- Data are extracted and processed behind the scenes:
- The Perseus Project (started in the 80s) automatically analyses grammar and vocabulary of ancient Greek texts and makes predictions about the function of a specific word in a specific passage
- So if the reader is not sure about a specific word (*nomoisi*) and clicks on it, a parallel window pops up to show several different dictionary entries all spelled *nomos*, gives statistical information, meanings, and links to further information.
- See the text and then the interpretations:



Herodotus, *The Histories*

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position.

This text is part of:

Greek and Roman Materials

Greek Prose

Greek Texts

Herodotus

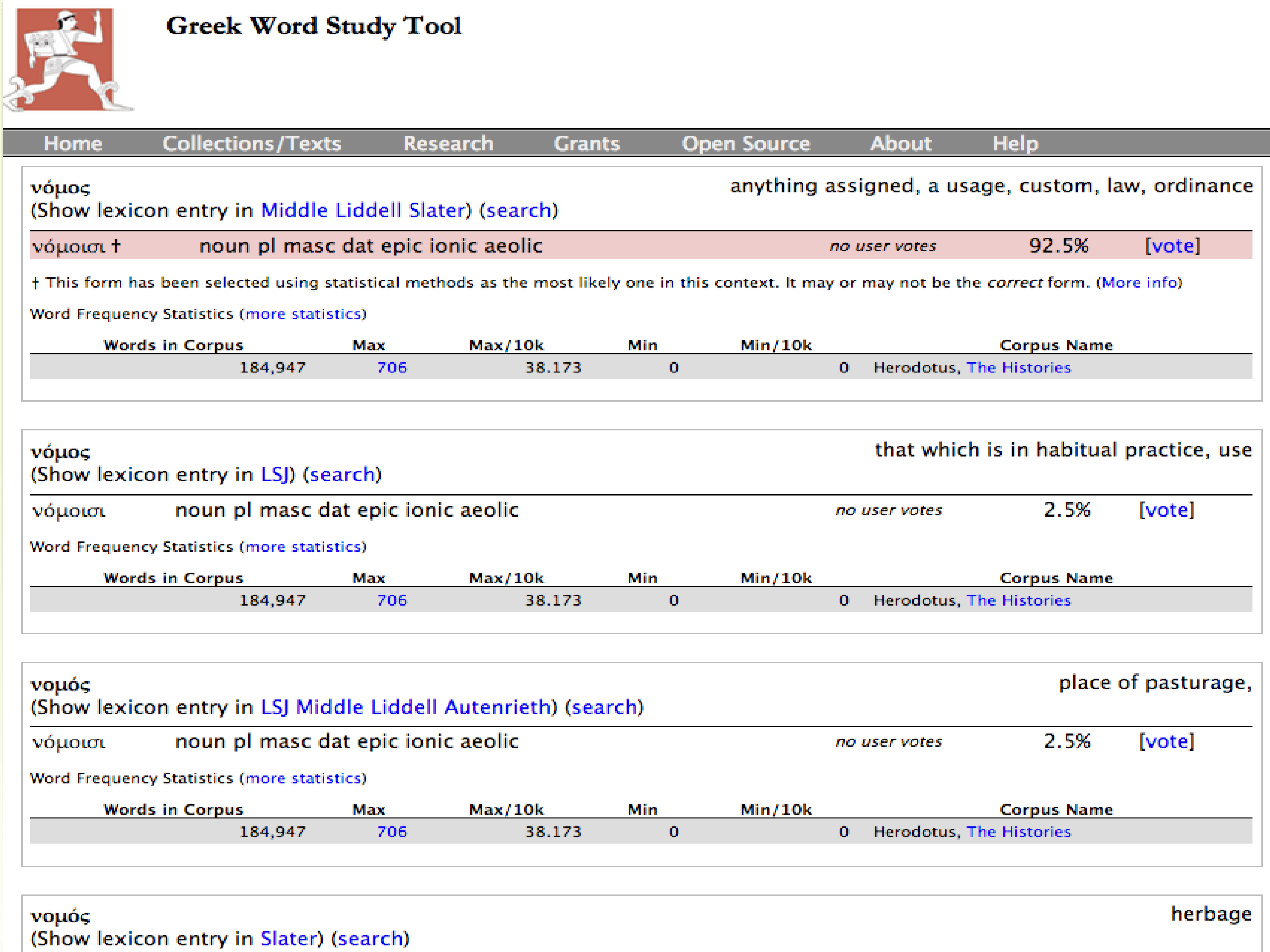
View text chunked by:

Hdt. 1.94.1

Click on a word to bring up parses, dictionary entries, and frequency statistics

Λυδοὶ δὲ νόμοισι μὲν παραπλησίοισι χρέωνται καὶ Ἕλληνές, χωρὶς ἢ ὅτι τὰ θήλεα τέκνα καταπορνεύουσι, πρῶτοι δὲ ἀνθρώπων τῶν ἡμεῖς ἴδμεν νόμισμα χρυσοῦ καὶ ἀργύρου κοψάμενοι ἐχρήσαντο, πρῶτοι δὲ καὶ κάπηλοι ἐγένοντο.

Herodotus, with an English translation by A. D. Godley. Cambridge. Harvard



Greek Word Study Tool

νόμος — anything assigned, a usage, custom, law, ordinance
(Show lexicon entry in Middle Liddell Slater) (search)

νόμοισι †	noun pl masc dat epic ionic aeolic	*no user votes*	92.5%	[vote]

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the *correct* form. (More info)

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
184,947	706	38.173	0	0	Herodotus, The Histories

νόμος — that which is in habitual practice, use
(Show lexicon entry in LSJ) (search)

νόμοισι	noun pl masc dat epic ionic aeolic	*no user votes*	2.5%	[vote]

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
184,947	706	38.173	0	0	Herodotus, The Histories

νομός — place of pasturage,
(Show lexicon entry in LSJ Middle Liddell Autenrieth) (search)

νόμοισι	noun pl masc dat epic ionic aeolic	*no user votes*	2.5%	[vote]

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
184,947	706	38.173	0	0	Herodotus, The Histories

νομός — herbage
(Show lexicon entry in Slater) (search)

Potential for research

- A critical mass of content now provides huge opportunities
  - Archives so large that there's no way to manually analyze them
  - A lot of this work starting in the past decade
  - Research needs rising, more demand in disciplines
- Efke Smit (May 2011): majority of scholarly publishers receive requests for mining less than 10 x per year; 5 for research purposes
- Publishers generally liberal in granting permissions; wish to understand this phenomenon and often go case by case
  - Reluctant to compete with their own products and services
- English is dominant language of scientific publication
- Don Swanson’s dream being realized

Happening now

- Newer, better software tools are being created to handle large corpora of content
- Biomed most advanced, along with corporate pharmaceutical and chemistry
- Growing number of users in government intelligence, business, social media, news, financial markets, legal arena
- Policy discussions at the national level, in the UK in particular, with the Hargreaves and JISC reports
- Lots of negative focus; given low, though rising demand, how did we so quickly get into a wrangle with publishers about TM?
- Automatic assumption that publishers have too many economic interests and won’t cooperate

Licenses

- Copyright/fair use status of text mining not clear
- Licenses can clarify what copyright does not
  - Licenses should permit automated queries for research and education – however it is done
  - Explicit language "text mining" or implicit language "for educational and research purposes"
- CDL have been seeking rights since 2008 and have secured some agreements using their standard language
- Overall, more and more library licenses have TM provision, but users not aware that library can help
- We will move from case by case to global rights and permissions – not so different after all from ILL

Next Steps

- Libraries can become more aware of campus needs and offer support/expertise
- Libraries can encourage publishers in licensing and researcher support.
- Topics span not just one publisher but cross publishers
  - There are numerous different publisher platforms – we need standards and standard content formats
  - How to do this? Cross-Rev, Cross-Mark, Cross-Check, Cross-Mine?
- Facilitating researchers' needs requires infrastructure development in various ways
- Collaboration is required across publishing, libraries, the research community
- Librarians can participate in facilitating such activities

Contact Information

- **Bernie Reilly, CRL**
  breilly@crl.edu
- **Virginia Kerr, CRL**
  vkerr@crl.edu
- **Ann Okerson, CRL**
  aokerson@crl.edu
- **David Magier, Princeton**
  dmagier@princeton.edu
- **Judson Dunham, Elsevier**
  j.dunham@elsevier.com
- **Ray Abruzzi, Cengage Learning**
  ray.abruzzi@cengage.com

Upcoming Events

CRL Webinar on News Collections in a Digital Age:

January 14, 2013

Annual Meeting of the CRL Council of Voting Members:

April 19, 2013

CRL Workshop at ALA Annual Meeting:

Global Resources Roundtable: "Beyond the Fold – Scholarly Access to News in the Digital Age"

June 27-28, 2013, Newberry Library, Chicago

For More Information

- Fill out our follow-up survey at

  http://www.surveymonkey.com/s/CRL_Text-Mining_Webinar_Nov_2012

- Check out other CRL presentations on our YouTube channel www.youtube.com/crldotedu

- Visit CRL website www.crl.edu

- Sign up for *CRL Connect:* www.crl.edu/connect

- Find CRL on Facebook and Twitter