# learning from nora:

distributed software development in the humanities

# www.noraproject.org

- Funded for 2004-2006, with $600K from the Andrew W. Mellon Foundation
- Participating researchers from Illinois (GSLIS, NCSA), Georgia (English), Maryland (MITH, HCIL, English), Virginia (IATH, CS, English), Alberta (Humanities Computing), Nebraska (English), McMaster (Multimedia)
- Fields of expertise include literary studies, library & information science, multimedia/design, computer science, computer engineering

# nora creation narrative

- “Tool-Time, or 'Haven't we been here before?': Ten Year in Humanities Computing” Delivered as part of "Transforming Disciplines: The Humanities and Computer Science," Saturday, January 18, 2003. Washington, DC.

# 2003

“We need (we still need) to demonstrate the usefulness of all the stuff we have digitized over the last decade and more--and usefulness not just in the form of increased access, but specifically, in what we can do with the stuff once we get it: what new questions we could ask, what old ones we could answer.”

# 2003

“We need to do this for two audiences: first, for colleagues in humanities departments who, while they admit that they are glad not to have to walk to the library to consult the library catalogue, can't really see that the digital library--assembled, inevitably, at the cost of other activities, services, and purchases--is really worth all that much. Second, we need to demonstrate this for the more general public, especially as it, and its values, gets represented in legislative priorities and state and federal funding.” -- “Tooltime”

# Nancy Ide, TSI: 1993

“Commercial software for text analysis and manipulation covers only a fraction of research needs, and it is often expensive and hard to adapt or extend to fit a particular research problem. Software developed by individual researchers and labs is often experimental and hard to get, hard to install, under-documented, and sometimes unreliable. Above all, most of this software is incompatible. As a result, it is not at all uncommon for researchers to develop tailor-made systems that replicate much of the functionality of other systems and in turn create programs that cannot be re-used by others, and so on in an endless software waste cycle.”

# 1996: CETH

- 1996: Center for Electronic Texts in the Humanities (Rutgers, Princeton) hosts a meeting responding, in part, to TSI's failure to effectively attract and mobilize volunteer labor.

# 1996: key features for next-gen text-analysis software

- modularity (a collection of relatively independent programs)
- professionality (should support serious research work)
- integration (modules should handle everything from data capture to analysis and presentation)
- portability (programs and data should be system-independent)

# Desire vs. Design

“The architectural group began, plausibly enough, by deciding to decide what it might mean to specify an architecture for the kind of system we had been talking about: what needs to be specified, and at what level of detail? This is, surely, a necessary first step. It would be nice to be able to report that after it, we had taken another one, but after we had reached something resembling agreement, it was time for lunch. Our dedication to the cause fought with our desire to eat; struggled; wavered; lost. We went to lunch.”

--Michael Sperberg-McQueen, from his 1996 trip report on Humanist, concerning the CETH meeting.

# 1998: ELTA

...a collaborative effort to encourage and support the development of software tools for the analysis, retrieval and manipulation of electronic texts. . . . We have organized Elta in response to continued interest and need for such software, most recently expressed at the birds-of-a-feather session at ALLC/ACH '98 in Debrecen. At this time Elta provides Web resources and an email list to support those interested in the Initiative's goals for promoting software development.

--Tom Horton, announcing ELTA on
Humanist in 1998

# 2003: What would it take to make something happen?

- **Consensus** on scholarly primitives and worthy problems among a “reasonable-sized group of researchers working with computational tools in humanities research”
- **Architectural specifications** (with as much as possible off the shelf)
- **Scale**: enough people working to create software while someone still wants to use it.

# 2003: What would it take to make something happen?

- **Management** to make sure that people actually are working together, are working on the same problems, are working toward a common goal.

- **Design and Testing** of these tools in conjunction with one another, in real research applications, with real researchers.

# 2003: Who is motivated to make it work?

- Foundations, agencies, and libraries that have made substantial investments in creating digital libraries are motivated to contribute funding, because they need to prove that the investment in digitizing--and especially in creating highly structured, high-quality digital collections--has been worth it.

# 2003-2004: nora pilot

**Mellon funds three meetings at UIUC:**

- Meeting 1: **Goals** (October 2003): big group
- Meeting 2: **Standards and Methods** (December 2003): smaller group
- Meeting 3: **Management (never happened).**
- Money re-budgeted for experimentation with D2K to prepare for a new proposal.

# nora's goal

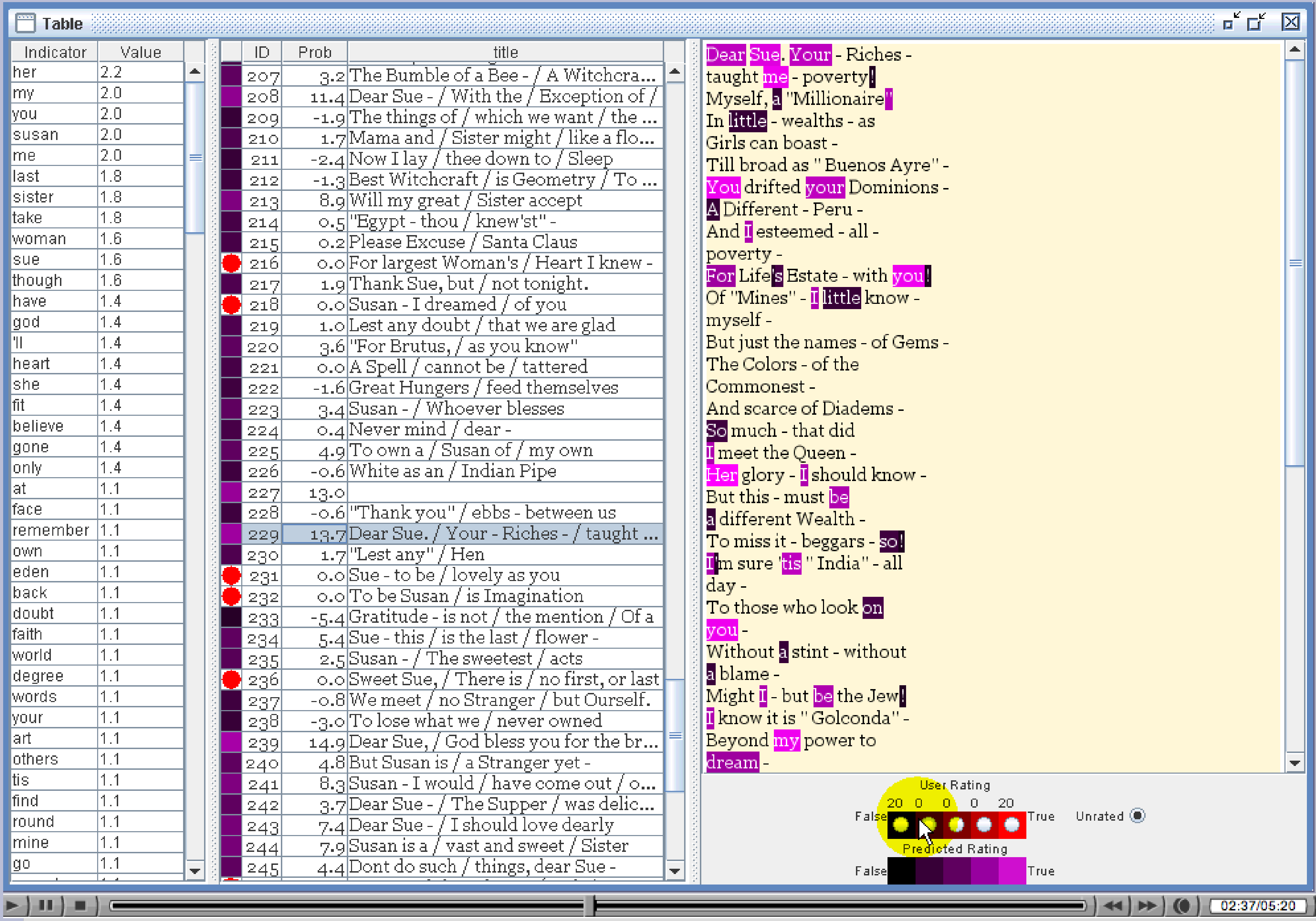
To produce software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries.

# 2004-2006+ : nora

- starts with 5 GB of 18th and 19th century British and American literature in SGML and XML contributed from about a dozen different libraries & projects

- version 1 in 2005, written in Java, using D2K and a postgres back-end designed by Steve Ramsay, and fed by GATE.

- version 2 in March 2007, written in Open Laszlo, using a proxy server between the interface and the datastore, which is now a combination of eXist and Lucene.  D2K still does the analytics.

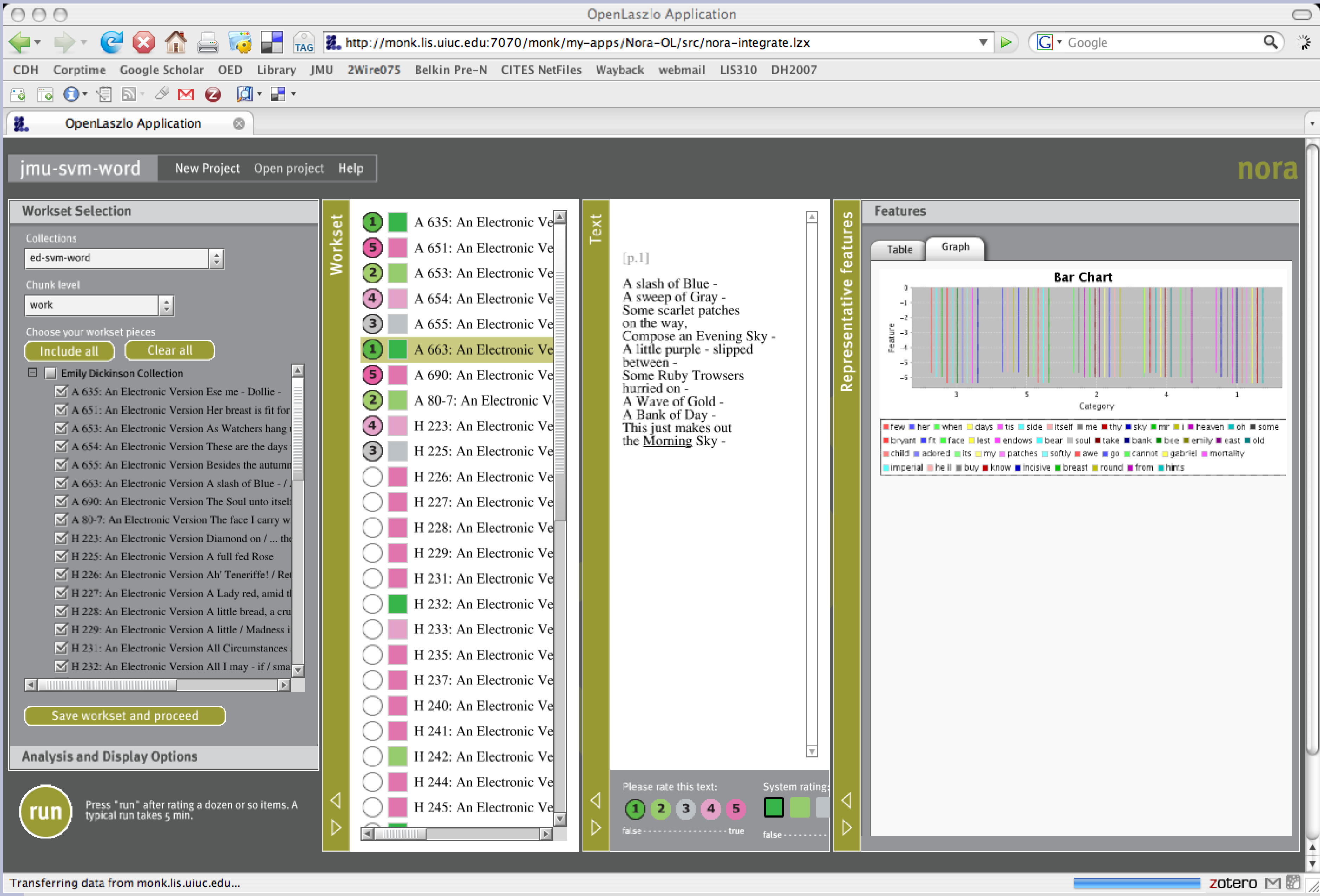


Table

| Indicator | Value |
|---|---|
| her | 2.2 |
| my | 2.0 |
| you | 2.0 |
| susan | 2.0 |
| me | 2.0 |
| last | 1.8 |
| sister | 1.8 |
| take | 1.8 |
| woman | 1.6 |
| sue | 1.6 |
| though | 1.6 |
| have | 1.4 |
| god | 1.4 |
| 'll | 1.4 |
| heart | 1.4 |
| she | 1.4 |
| fit | 1.4 |
| believe | 1.4 |
| gone | 1.4 |
| only | 1.4 |
| at | 1.1 |
| face | 1.1 |
| remember | 1.1 |
| own | 1.1 |
| eden | 1.1 |
| back | 1.1 |
| doubt | 1.1 |
| faith | 1.1 |
| world | 1.1 |
| degree | 1.1 |
| words | 1.1 |
| your | 1.1 |
| art | 1.1 |
| others | 1.1 |
| tis | 1.1 |
| find | 1.1 |
| round | 1.1 |
| mine | 1.1 |
| go | 1.1 |

| | ID | Prob | title |
|---|---|---|---|
| ■ | 207 | 3.2 | The Bumble of a Bee - / A Witchcra... |
| ■ | 208 | 11.4 | Dear Sue - / With the / Exception of / |
| ■ | 209 | -1.9 | The things of / which we want / the ... |
| ■ | 210 | 1.7 | Mama and / Sister might / like a flo... |
| ■ | 211 | -2.4 | Now I lay / thee down to / Sleep |
| ■ | 212 | -1.3 | Best Witchcraft / is Geometry / To ... |
| ■ | 213 | 8.9 | Will my great / Sister accept |
| ■ | 214 | 0.5 | "Egypt - thou / knew'st" - |
| ■ | 215 | 0.2 | Please Excuse / Santa Claus |
| ● | 216 | 0.0 | For largest Woman's / Heart I knew - |
| ■ | 217 | 1.9 | Thank Sue, but / not tonight. |
| ● | 218 | 0.0 | Susan - I dreamed / of you |
| ■ | 219 | 1.0 | Lest any doubt / that we are glad |
| ■ | 220 | 3.6 | "For Brutus, / as you know" |
| ■ | 221 | 0.0 | A Spell / cannot be / tattered |
| ■ | 222 | -1.6 | Great Hungers / feed themselves |
| ■ | 223 | 3.4 | Susan - / Whoever blesses |
| ■ | 224 | 0.4 | Never mind / dear - |
| ■ | 225 | 4.9 | To own a / Susan of / my own |
| ■ | 226 | -0.6 | White as an / Indian Pipe |
| ■ | 227 | 13.0 | |
| ■ | 228 | -0.6 | "Thank you" / ebbs - between us |
| ■ | 229 | 13.7 | Dear Sue. / Your - Riches - / taught ... |
| ■ | 230 | 1.7 | "Lest any" / Hen |
| ● | 231 | 0.0 | Sue - to be / lovely as you |
| ● | 232 | 0.0 | To be Susan / is Imagination |
| ■ | 233 | -5.4 | Gratitude - is not / the mention / Of a |
| ■ | 234 | 5.4 | Sue - this / is the last / flower - |
| ■ | 235 | 2.5 | Susan - / The sweetest / acts |
| ● | 236 | 0.0 | Sweet Sue, / There is / no first, or last |
| ■ | 237 | -0.8 | We meet / no Stranger / but Ourself. |
| ■ | 238 | -3.0 | To lose what we / never owned |
| ■ | 239 | 14.9 | Dear Sue, / God bless you for the br... |
| ■ | 240 | 4.8 | But Susan is / a Stranger yet - |
| ■ | 241 | 8.3 | Susan - I would / have come out / o... |
| ■ | 242 | 3.7 | Dear Sue - / The Supper / was delic... |
| ■ | 243 | 7.4 | Dear Sue - / I should love dearly |
| ■ | 244 | 7.9 | Susan is a / vast and sweet / Sister |
| ■ | 245 | 4.4 | Dont do such / things, dear Sue - |

Dear Sue. Your - Riches -
taught me - poverty!
Myself, a "Millionaire"
In little - wealths - as
Girls can boast -
Till broad as " Buenos Ayre" -
You drifted your Dominions -
A Different - Peru -
And I esteemed - all -
poverty -
For Life's Estate - with you!
Of "Mines" - I little know -
myself -
But just the names - of Gems -
The Colors - of the
Commonest -
And scarce of Diadems -
So much - that did
I meet the Queen -
Her glory - I should know -
But this - must be
a different Wealth -
To miss it - beggars - so!
I'm sure 'tis " India" - all
day -
To those who look on
you -
Without a stint - without
a blame -
Might I - but be the Jew!
I know it is " Golconda" -
Beyond my power to
dream -



OpenLaszlo Application

http://monk.lis.uiuc.edu:7070/monk/my-apps/Nora-OL/src/nora-integrate.lzx

Google
CDH Corptime Google Scholar OED Library JMU 2Wire075 Belkin Pre-N CITES NetFiles Wayback webmail LIS310 DH2007
OpenLaszlo Application
jmu-svm-word
New Project
Open project
Help
nora

Workset Selection

Collections: ed-svm-word

Chunk level: work

Choose your workset pieces

□ Emily Dickinson Collection
- ☑ A 635: An Electronic Version Ese me - Dollie -
- ☑ A 651: An Electronic Version Her breast is fit for
- ☑ A 653: An Electronic Version As Watchers hang u
- ☑ A 654: An Electronic Version These are the days
- ☑ A 655: An Electronic Version Besides the autumn
- ☑ A 663: An Electronic Version A slash of Blue - / A
- ☑ A 690: An Electronic Version The Soul unto itsel
- ☑ A 80-7: An Electronic Version The face I carry w
- ☑ H 223: An Electronic Version Diamond on / ... th
- ☑ H 225: An Electronic Version A full fed Rose
- ☑ H 226: An Electronic Version Ah' Teneriffe! / Ret
- ☑ H 227: An Electronic Version A Lady red, amid t
- ☑ H 228: An Electronic Version A little bread, a cru
- ☑ H 229: An Electronic Version A little / Madness i
- ☑ H 231: An Electronic Version All Circumstances
- ☑ H 232: An Electronic Version All I may - if / sma

Analysis and Display Options

Press "run" after rating a dozen or so items. A typical run takes 5 min.

- 1 A 635: An Electronic Ve
- 5 A 651: An Electronic Ve
- 2 A 653: An Electronic Ve
- 4 A 654: An Electronic Ve
- 3 A 655: An Electronic Ve
- 1 A 663: An Electronic Ve
- 5 A 690: An Electronic Ve
- 2 A 80-7: An Electronic V
- 4 H 223: An Electronic Ve
- 3 H 225: An Electronic Ve
- ○ H 226: An Electronic Ve
- ○ H 227: An Electronic Ve
- ○ H 228: An Electronic Ve
- ○ H 229: An Electronic Ve
- ○ H 231: An Electronic Ve
- ○ H 232: An Electronic Ve
- ○ H 233: An Electronic Ve
- ○ H 235: An Electronic Ve
- ○ H 237: An Electronic Ve
- ○ H 240: An Electronic Ve
- ○ H 241: An Electronic Ve
- ○ H 242: An Electronic Ve
- ○ H 244: An Electronic Ve
- ○ H 245: An Electronic Ve

[p.1]

A slash of Blue -
A sweep of Gray -
Some scarlet patches
on the way,
Compose an Evening Sky -
A little purple - slipped
between -
Some Ruby Trowsers
hurried on -
A Wave of Gold -
A Bank of Day -
This just makes out
the Morning Sky -

Please rate this text: 1 2 3 4 5

false - - - - - - - - - - - - - - - - true

System rating:

false - - - - - - - -

Transferring data from monk.lis.uiuc.edu...

# 2007: lessons learned

“In two years, we have made good progress toward [nora's original] goal, but we have also encountered some significant challenges along the way.  In retrospect, some of these—particularly with respect to project management and dependencies created by the design of nora's software architecture—could have been avoided.”

--final report on nora to the Mellon Foundation

# 2005: In Praise of Pattern

“The exploration of pattern may be usefully regarded as the strongest point of intersection between the computational strictures of text analysis and the open-ended interpretive landscape of literary studies....”

# 2005: In Praise of Pattern

"...Seeing computational analysis in literary studies as a quest for interpretations inspired by pattern can, moreover, lead to a change in the perception of text analysis among more mainstream literary critics by moving the hermeneutical justification of the activity away from the denotative realm of science and toward the more broadly rhetorical and exegetical practices of the humanities."

--Steve Ramsay, *Text Technology*, 2005.

# 2006: In Praise of Readers

“The point ... is not to save the reader from reading the individual texts or from making an independent judgment of each document's characteristics; rather, the point is for nora to learn from the reader's holistic impression of the text and then, having done so, to show the reader what evidence correlates with these impressions...”

# 2006: In Praise of Readers

"...Dickinson uses a small vocabulary in her poems--a few thousand words--but even the keenest human reader cannot reliably keep track of the frequency with which each of those words is used across even a small set of documents. This is what nora can do. What nora cannot do, of course, is explain the results. That remains the task of the reader."

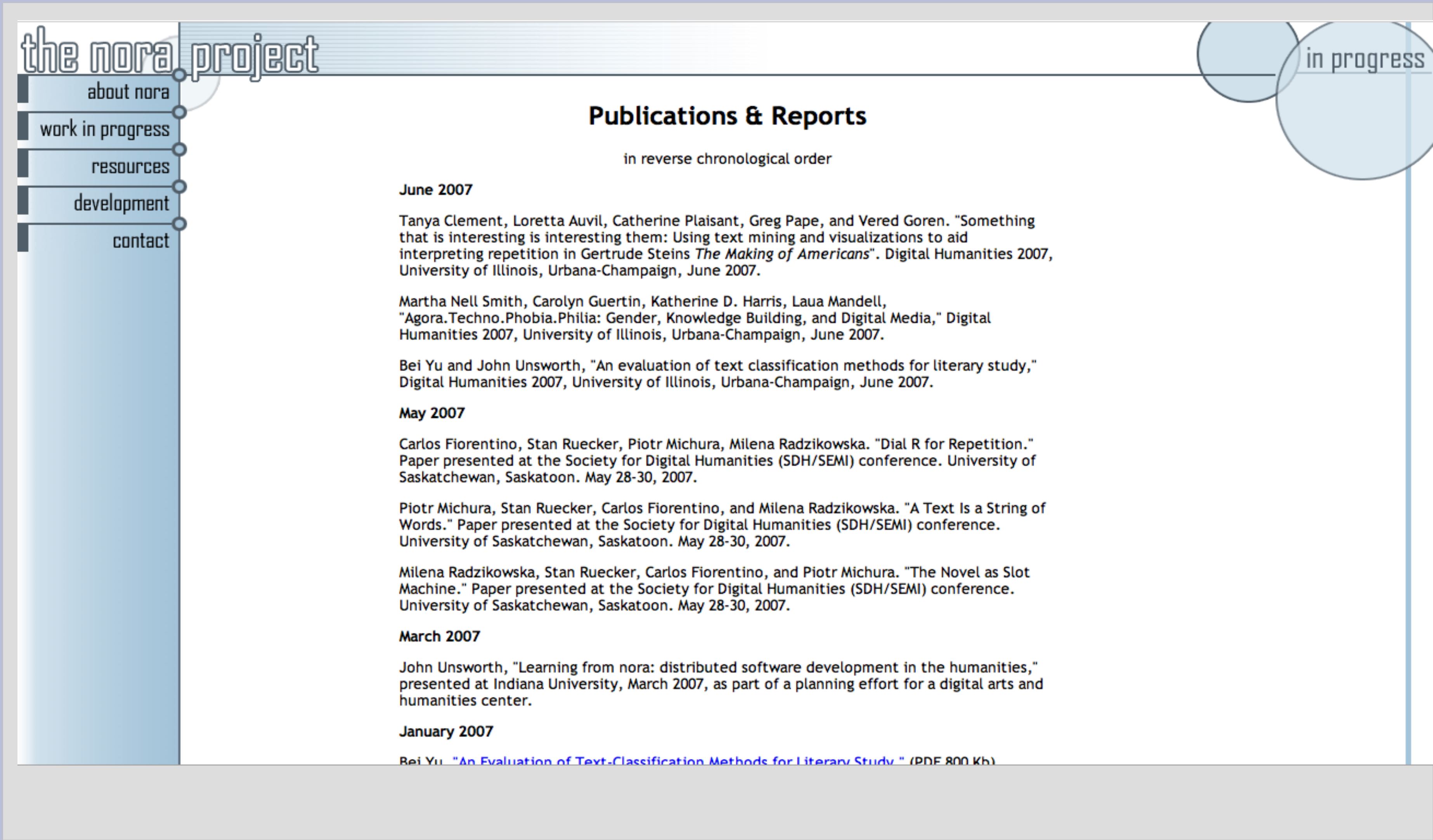
-- nora help docs

# Provocation vs. Prediction

“Provocation, in the context of data mining—where there is typically an expectation of ground truth and verifiable results—is a non-trivial intervention....What we've done with nora represents an important applied extension of contributions by people like Jerry McGann, Johanna Drucker, and Willard McCarty, who have theorized the role of deformation, provocation, modeling, and play in the humanities.”

-- nora final report to Mellon

# 40 publications/presentations from 18 authors at 8 institutions



## Publications & Reports

in reverse chronological order

**June 2007**

Tanya Clement, Loretta Auvil, Catherine Plaisant, Greg Pape, and Vered Goren. "Something that is interesting is interesting them: Using text mining and visualizations to aid interpreting repetition in Gertrude Steins *The Making of Americans*". Digital Humanities 2007, University of Illinois, Urbana-Champaign, June 2007.

Martha Nell Smith, Carolyn Guertin, Katherine D. Harris, Laua Mandell, "Agora.Techno.Phobia.Philia: Gender, Knowledge Building, and Digital Media," Digital Humanities 2007, University of Illinois, Urbana-Champaign, June 2007.

Bei Yu and John Unsworth, "An evaluation of text classification methods for literary study," Digital Humanities 2007, University of Illinois, Urbana-Champaign, June 2007.

**May 2007**

Carlos Fiorentino, Stan Ruecker, Piotr Michura, Milena Radzikowska. "Dial R for Repetition." Paper presented at the Society for Digital Humanities (SDH/SEMI) conference. University of Saskatchewan, Saskatoon. May 28-30, 2007.

Piotr Michura, Stan Ruecker, Carlos Fiorentino, and Milena Radzikowska. "A Text Is a String of Words." Paper presented at the Society for Digital Humanities (SDH/SEMI) conference. University of Saskatchewan, Saskatoon. May 28-30, 2007.

Milena Radzikowska, Stan Ruecker, Carlos Fiorentino, and Piotr Michura. "The Novel as Slot Machine." Paper presented at the Society for Digital Humanities (SDH/SEMI) conference. University of Saskatchewan, Saskatoon. May 28-30, 2007.

**March 2007**

John Unsworth, "Learning from nora: distributed software development in the humanities," presented at Indiana University, March 2007, as part of a planning effort for a digital arts and humanities center.

**January 2007**

Bei Yu, "An Evaluation of Text-Classification Methods for Literary Study," (PDF 800 Kb)

# management 101

It is necessary to subdivide large distributed projects into functional sub-units, each with their own leadership and goals, with regular (and documented) conference calls or face-to-face meetings, and with a structure for reporting up to a coordinating group that parcels out tasks to sub-groups and keeps track of whether those tasks are getting done, and also pays attention to when a lack of progress in one part of the project is impeding progress in another part.

# management 102

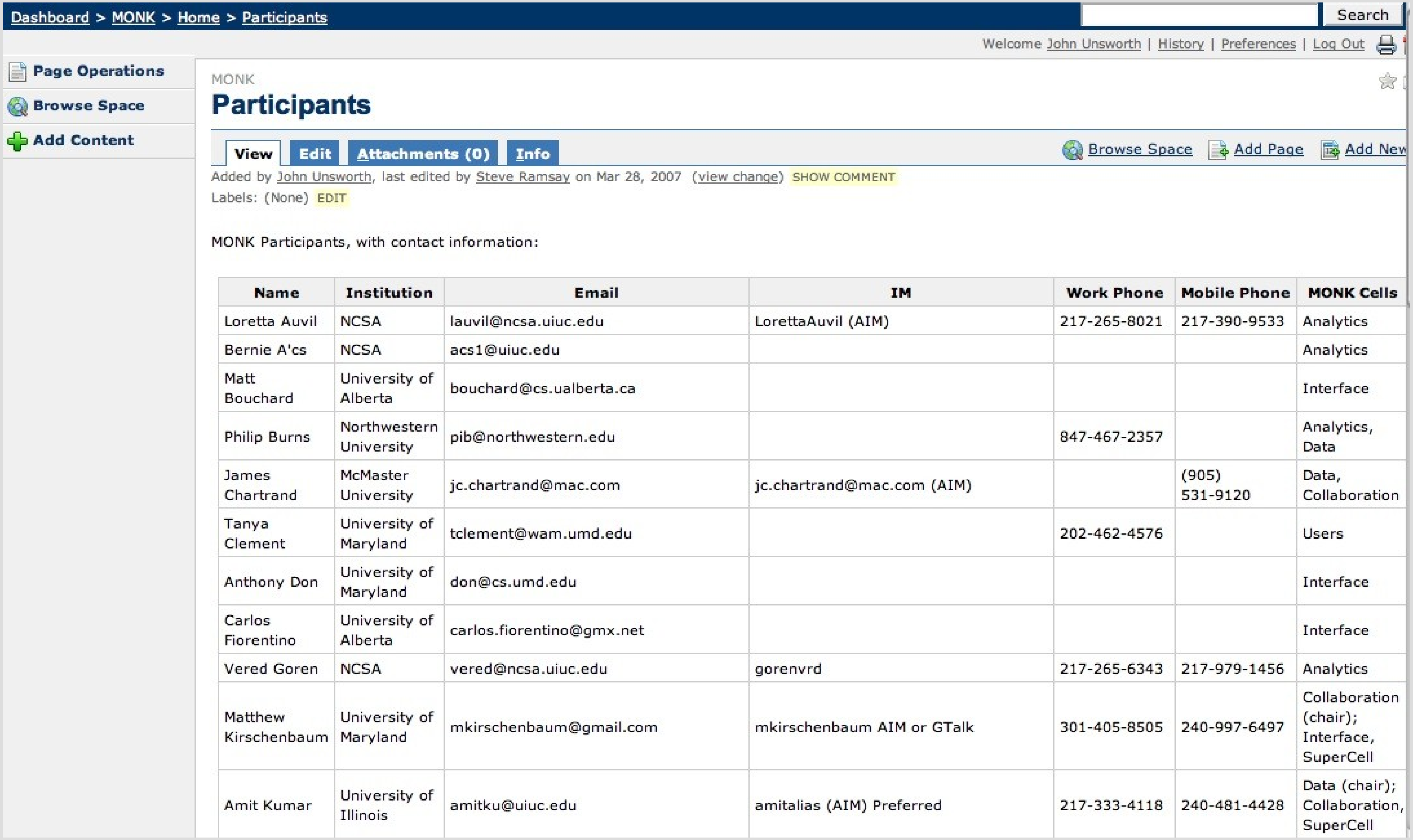
Regular face-to-face meetings, including face-to-face all-hands meetings, are critically important to maintaining momentum in a multi-participant, multi-institutional project.  Without effective sub-division, the nora group was too large, and had too many different agendas, to function effectively in a common conference call: people ended up frustrated that the issues most important to them had not been discussed enough.

# The Mythical Man-Month

“In most projects, the first system built is barely usable. It may be too slow, too big, awkward to use, or all three. There is no alternative but to start again, smarting but smarter, and build a redesigned version in which these problems are solved. . . . Where a new system concept or a new technology is used, one has to build a system to throw away, for even the best planning is not so omniscient as to get it right the first time.”

--Frederick Brooks, The Mythical Man-Month

# MONK participants



MONK

## Participants

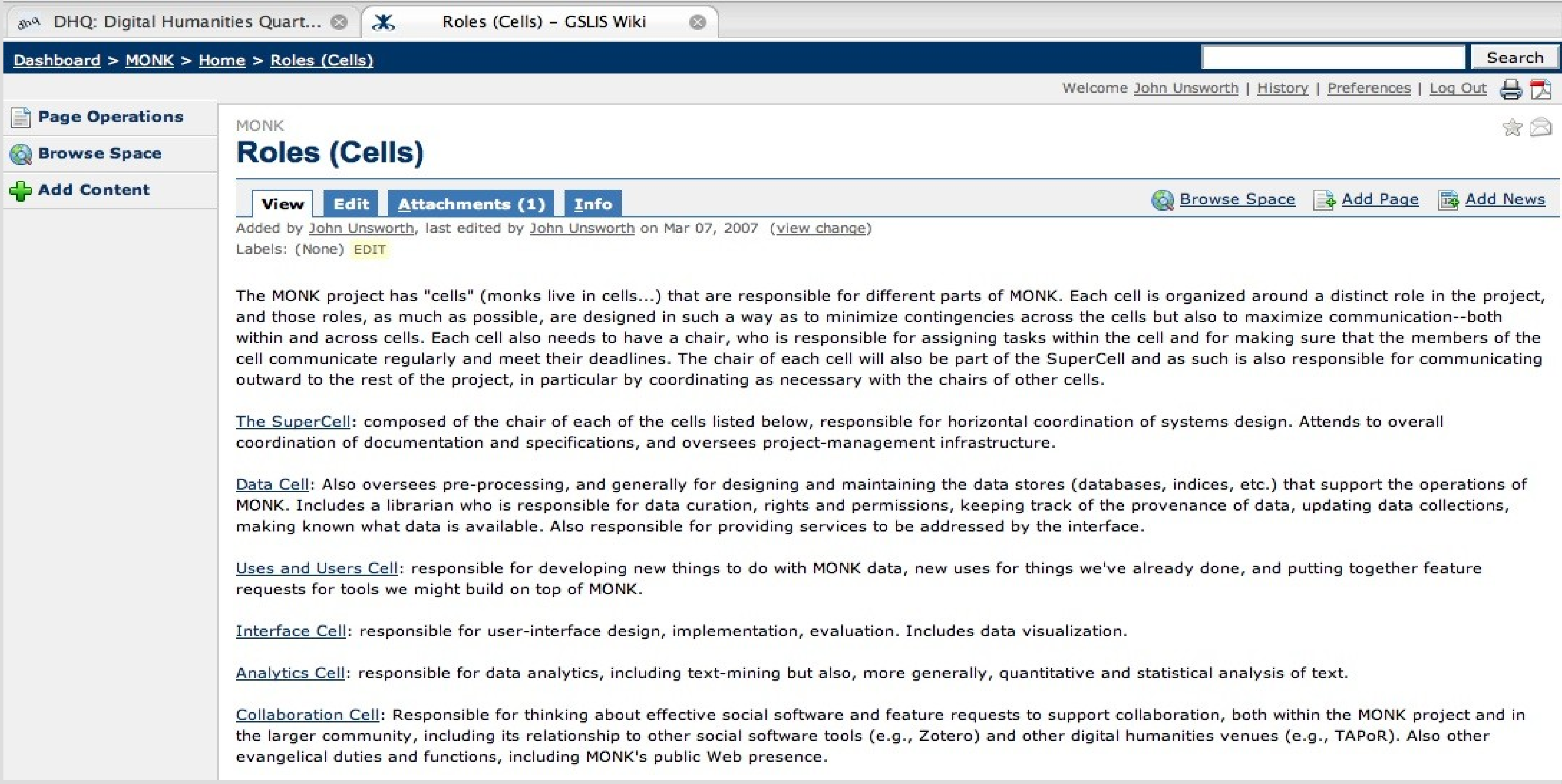
Added by John Unsworth, last edited by Steve Ramsay on Mar 28, 2007 (view change)

Labels: (None)

MONK Participants, with contact information:

| Name | Institution | Email | IM | Work Phone | Mobile Phone | MONK Cells |
|---|---|---|---|---|---|---|
| Loretta Auvil | NCSA | lauvil@ncsa.uiuc.edu | LorettaAuvil (AIM) | 217-265-8021 | 217-390-9533 | Analytics |
| Bernie A'cs | NCSA | acs1@uiuc.edu | | | | Analytics |
| Matt Bouchard | University of Alberta | bouchard@cs.ualberta.ca | | | | Interface |
| Philip Burns | Northwestern University | pib@northwestern.edu | | 847-467-2357 | | Analytics, Data |
| James Chartrand | McMaster University | jc.chartrand@mac.com | jc.chartrand@mac.com (AIM) | | (905) 531-9120 | Data, Collaboration |
| Tanya Clement | University of Maryland | tclement@wam.umd.edu | | 202-462-4576 | | Users |
| Anthony Don | University of Maryland | don@cs.umd.edu | | | | Interface |
| Carlos Fiorentino | University of Alberta | carlos.fiorentino@gmx.net | | | | Interface |
| Vered Goren | NCSA | vered@ncsa.uiuc.edu | gorenvrd | 217-265-6343 | 217-979-1456 | Analytics |
| Matthew Kirschenbaum | University of Maryland | mkirschenbaum@gmail.com | mkirschenbaum AIM or GTalk | 301-405-8505 | 240-997-6497 | Collaboration (chair); Interface, SuperCell |
| Amit Kumar | University of Illinois | amitku@uiuc.edu | amitalias (AIM) Preferred | 217-333-4118 | 240-481-4428 | Data (chair); Collaboration, SuperCell |

# MONK roles


DHQ: Digital Humanities Quart...
Roles (Cells) – GSLIS Wiki
Dashboard > MONK > Home > Roles (Cells)
Search
Welcome John Unsworth | History | Preferences | Log Out
Page Operations
Browse Space
Add Content
View
Edit
Attachments (1)
Info
Browse Space
Add Page
Add News

MONK

## Roles (Cells)

Added by John Unsworth, last edited by John Unsworth on Mar 07, 2007 (view change)

Labels: (None) EDIT

The MONK project has "cells" (monks live in cells...) that are responsible for different parts of MONK. Each cell is organized around a distinct role in the project, and those roles, as much as possible, are designed in such a way as to minimize contingencies across the cells but also to maximize communication--both within and across cells. Each cell also needs to have a chair, who is responsible for assigning tasks within the cell and for making sure that the members of the cell communicate regularly and meet their deadlines. The chair of each cell will also be part of the SuperCell and as such is also responsible for communicating outward to the rest of the project, in particular by coordinating as necessary with the chairs of other cells.

The SuperCell: composed of the chair of each of the cells listed below, responsible for horizontal coordination of systems design. Attends to overall coordination of documentation and specifications, and oversees project-management infrastructure.

Data Cell: Also oversees pre-processing, and generally for designing and maintaining the data stores (databases, indices, etc.) that support the operations of MONK. Includes a librarian who is responsible for data curation, rights and permissions, keeping track of the provenance of data, updating data collections, making known what data is available. Also responsible for providing services to be addressed by the interface.

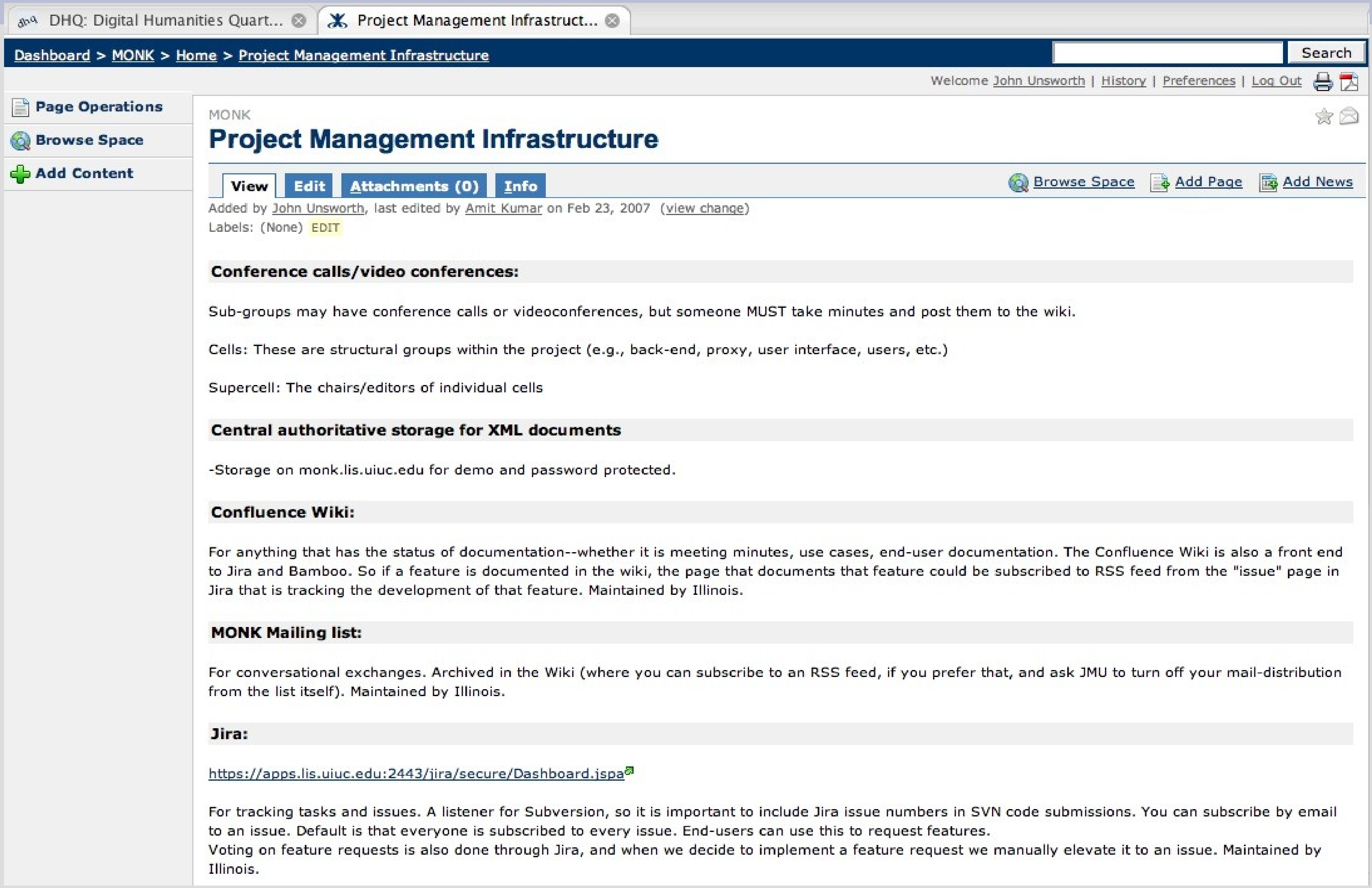
Uses and Users Cell: responsible for developing new things to do with MONK data, new uses for things we've already done, and putting together feature requests for tools we might build on top of MONK.

Interface Cell: responsible for user-interface design, implementation, evaluation. Includes data visualization.

Analytics Cell: responsible for data analytics, including text-mining but also, more generally, quantitative and statistical analysis of text.

Collaboration Cell: Responsible for thinking about effective social software and feature requests to support collaboration, both within the MONK project and in the larger community, including its relationship to other social software tools (e.g., Zotero) and other digital humanities venues (e.g., TAPoR). Also other evangelical duties and functions, including MONK's public Web presence.
]

# MONK management


DHQ: Digital Humanities Quart...
Project Management Infrastruct...
Dashboard > MONK > Home > Project Management Infrastructure
Search
Welcome John Unsworth | History | Preferences | Log Out
Page Operations
Browse Space
Add Content
View
Edit
Attachments (0)
Info
Browse Space
Add Page
Add News

MONK

## Project Management Infrastructure

Added by John Unsworth, last edited by Amit Kumar on Feb 23, 2007 (view change)

Labels: (None) EDIT

### Conference calls/video conferences:

Sub-groups may have conference calls or videoconferences, but someone MUST take minutes and post them to the wiki.

Cells: These are structural groups within the project (e.g., back-end, proxy, user interface, users, etc.)

Supercell: The chairs/editors of individual cells

### Central authoritative storage for XML documents

-Storage on monk.lis.uiuc.edu for demo and password protected.

### Confluence Wiki:

For anything that has the status of documentation--whether it is meeting minutes, use cases, end-user documentation. The Confluence Wiki is also a front end to Jira and Bamboo. So if a feature is documented in the wiki, the page that documents that feature could be subscribed to RSS feed from the "issue" page in Jira that is tracking the development of that feature. Maintained by Illinois.

### MONK Mailing list:

For conversational exchanges. Archived in the Wiki (where you can subscribe to an RSS feed, if you prefer that, and ask JMU to turn off your mail-distribution from the list itself). Maintained by Illinois.

### Jira:

https://apps.lis.uiuc.edu:2443/jira/secure/Dashboard.jspa

For tracking tasks and issues. A listener for Subversion, so it is important to include Jira issue numbers in SVN code submissions. You can subscribe by email to an issue. Default is that everyone is subscribed to every issue. End-users can use this to request features.
Voting on feature requests is also done through Jira, and when we decide to implement a feature request we manually elevate it to an issue. Maintained by Illinois.