The Role of Academic Research Libraries in the Digital Data Universe

Rick Luce, Emory University

Librarians and eScience: Focusing towards 20/20

CIC, Purdue – May 13, 2008

Environmental Scan → where our customers live

- Dawn of technology enabled social tools

  New ways to connect and collaborate

- Rise of eResearch, digital humanities, eScience

  New ways to work: digital scholarship, eResearch

- Collective problems require collection action, which requires a shared vision

  Concept of 'cloud collaboration'

  Leadership imperatives

Context: Changing User Expectations

- Changing user expectations
  - 70 percent of Emory students own iPods
  - Another 20 percent have other portable devices
    = mini digital libraries

  - A pattern: open access, open science, open source, open data…
  - Connectivity, collaboration, social networking
    - Blogging, Podcasting and Vodcasting

"*The future ain't what it used to be*" -- Yogi Berra

From Sharing to Cooperation to Collective Action

As transistors led to computers and fundamental societal changes, so will

social media and Web 2.0 / 3.0

- Online social networks enable collaborative groups to form regardless of geography —groups larger, more distributed than at any other time in history

Group Activities Enabled by Social Media

Three levels:

1. Sharing via social tools: del.icio.us, Flickr, Slideshare
   - After 9/11, a Middle Eastern history prof's blog became a resource for reporters covering battles in Afghanistan & Iraq
2. Collaboration - e.g., using Linux or Wikipedia
3. Collective action - groups form to pursue a larger purpose and use social tools, ranging from Google or Yahoo! groups to free online social networks (Ning) to share news and tips, recruit others, support each other and remain unified.

See: Clay Shirky - *Here Comes Everybody*

Evolving Science and the Emergence of eScience

- **Thousand years** ago:
  Empirical science **-** *experimentation describing natural phenomena*

- Last few **hundred years**:

Theoretical science *using models, generalizations*

- Last **few decades**:

Computational science - *simulating complex phenomena*

- **Today**:

Data exploration / data science **-** *synthesizing theory, experimentation and computation with advanced data management and statistics → new algorithms*

Changing Nature of Research in Science

eScience is characterized as computationally intense science employing grid computing technology to analyze very large data collections - John Taylor

Instead of spending six months doing an experiment which you can then understand in an afternoon when you're done, you can do an experiment in an afternoon and it takes you six months to figure out what you've got

Chronicle: *Learning to Swim in the Rising Tide of Scientific Data* - 6/29/01

Open Access to Scientific Research

Access to scientific research is an obligation of the scientific community, both for the advancement of science as well as the responsibility to account for public investments

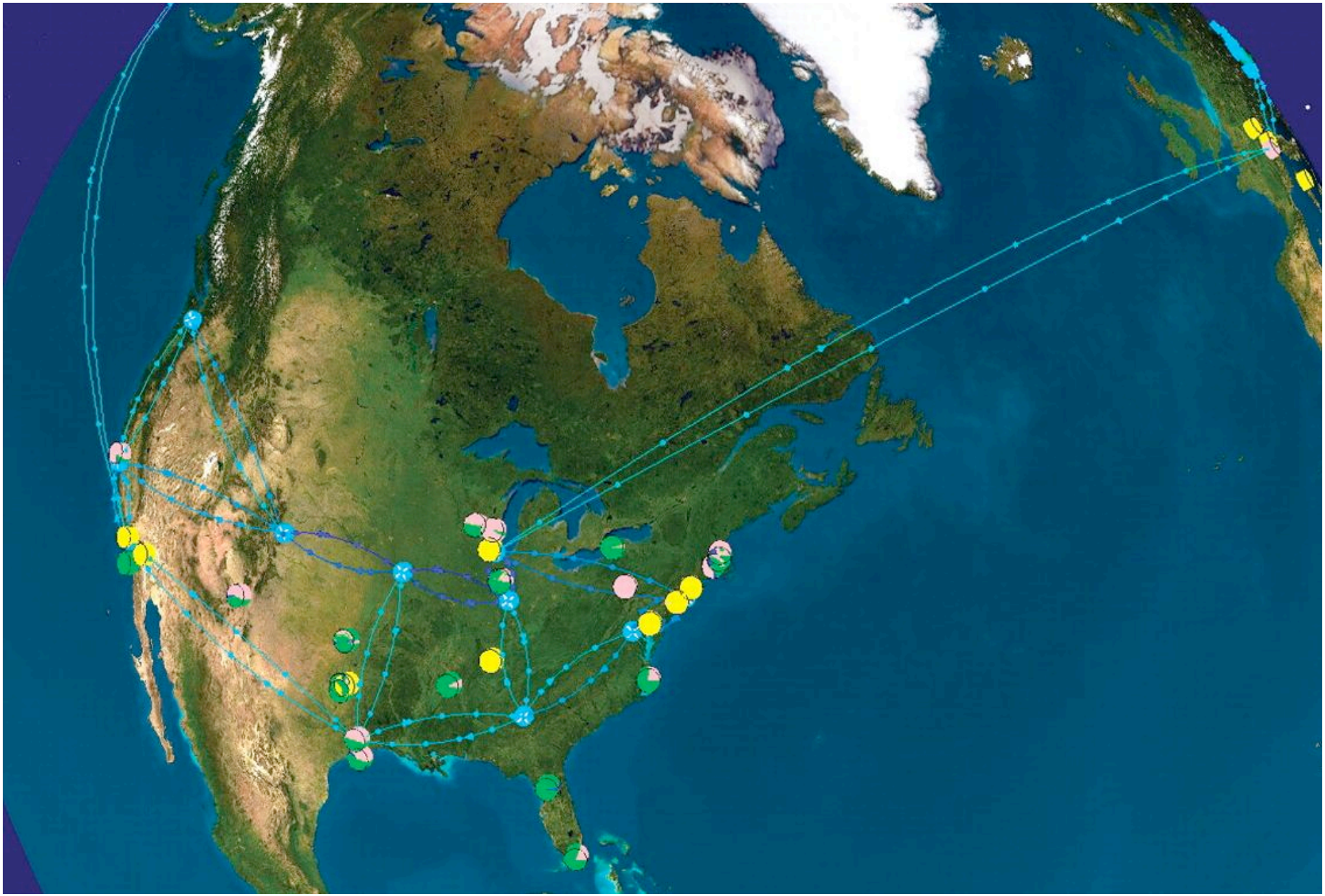
New requirements arise to support eScience –

- *Enabled by cyberinfrastructure, used to create new knowledge, we can see new approaches to scientific discovery linked together in a global network with far greater interactivity and broader collaboration**

- *Such collaboratories, grid or virtual science communities, etc., are dependent upon open access for the creation, dissemination and preservation of knowledge**

**Atkins NSF report: Revolutionizing Science and Engineering through Cyberinfrastructure

Open Science Grid: links storage and computing resources at more than 30 sites across the USA



"Those who do not learn from the future are destined to make mistakes in it."

Rick Luce, CIC, 5/12/08

Looking Ahead: Perspective from Science

Science is changing the process of how we know things - and the foundation of our culture and knowledge

- Hypothesis search and deep real time simulations drive data collection and information manipulation

Distributed instrumentation and experiments will yield smart-mob, hive mind science operating fast, cheap, and out of control

- Triple blind experiments emerge through massive non-invasive statistical data. No one realizes experiment is going on till much later…
- Negative results have positive value
  - see: *Journal of Negative Results* in Biomedicine

Emerging new ways of knowing in science

- **Evolutionary search**: combinatorial exploration of variations derived from the best of a previous generation of good results. Best results are mutated and bred for better results
- **Multiple Hypothesis Matrix**: matrix of many scenarios are proposed and managed simultaneously
- **Adaptive Real Time Experiments**: real time result evaluation and modification of large scale experiments. Analysis happens in parallel with collection and design of the test is shifted on the fly
- **Wiki-science**: experiments involving thousands of investigators collaborating on a 'paper' which is ongoing and never finished
  - Tools for tracking credit and contributions are vital
- **Zillionics:** ubiquitous always-on sensors in bodies and environments
- **Intelligent bio-machines**: putting nanobots into our bloodstream

Hot Data publishing Issues

Data publishing on the grid:

- **Data integration** - tying together data from various sources
- **Annotation** - adding comments & observations to existing data, becoming a new form of communication
- **Provenance** - where did this data come from?
- **Exporting/publishing** in agreed formats - to other programs & people
- **Security** - Specifying/enforcing read/write access to *parts* of your data

A Serial Model of Research and Scholarship

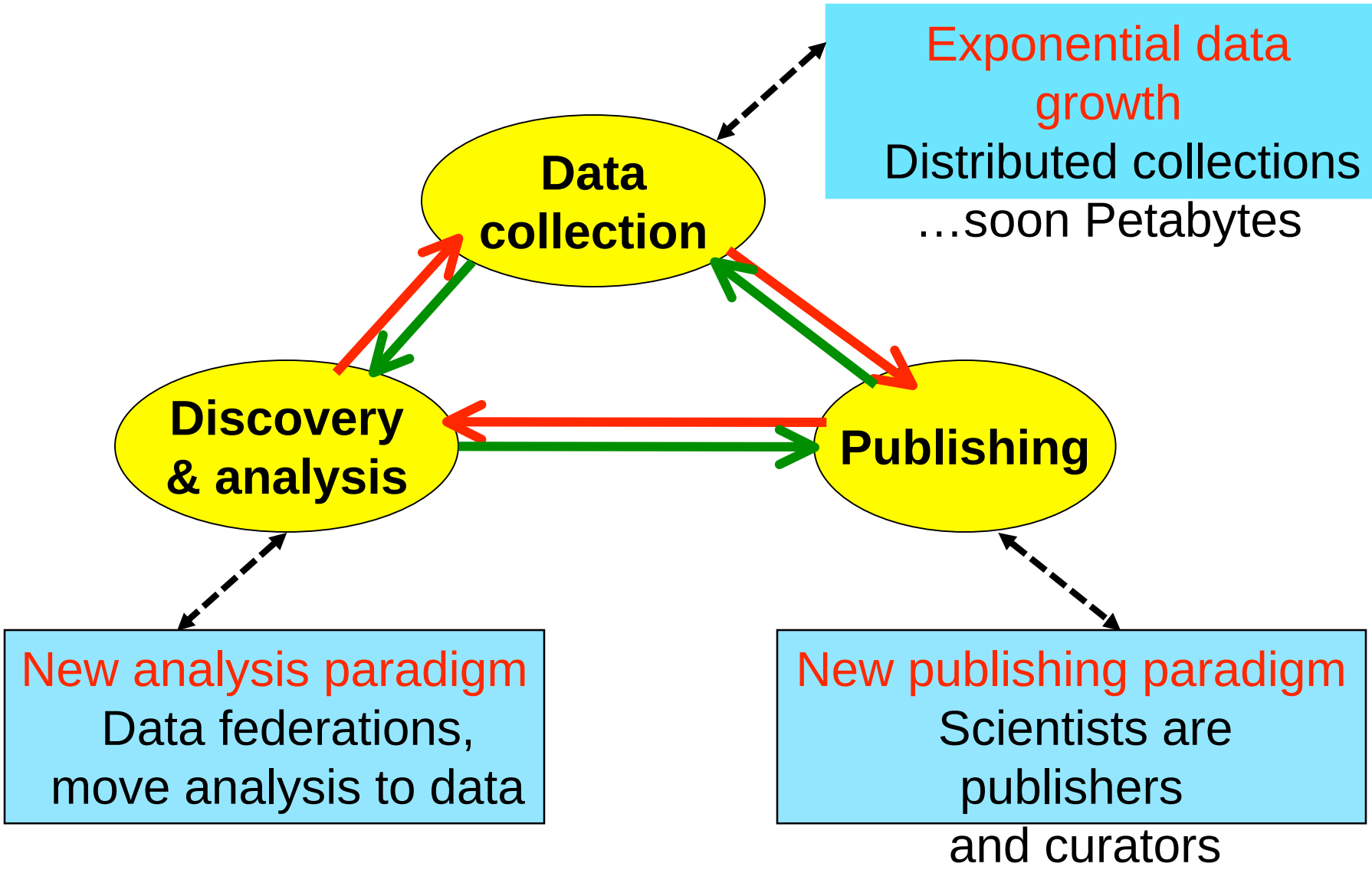
(the model many of us grew up with)

Managing a Data Deluge

Collecting data

- Very extended distribution of data sets
  - Most datasets are small, and manually maintained - spreadsheets, servers
  - Most bytes today are collected via electronic sensors
- Leads to living in an exponential world
  - Caused by the emergence of generations of inexpensive sensors + computing
  - Projects last ~ 2-5 years, data sent upwards only at project end
  - Data doubles every year - data will **never** be centralized
- PI's taking more responsibility on projects
  - Becoming publishers and curators (Data Publishing)
  - Data resides with projects - analyses must be close to the data

The New Parallel Model



New Roles for Research Libraries

An opportunity for the uniquely placed digital science library to connect disciplines and newly emerging fields

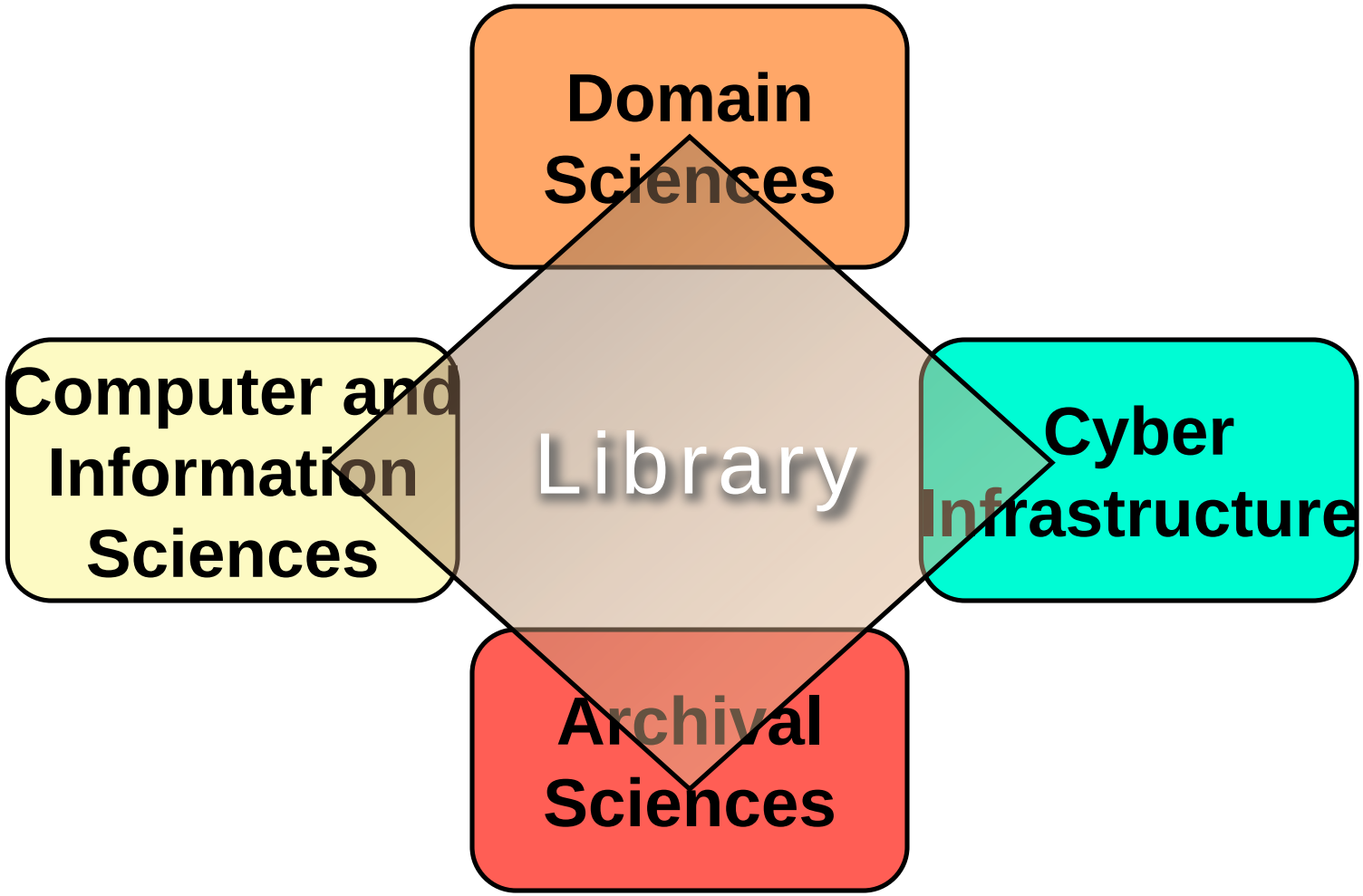
- Changing landscape for publications and scholarly communication
- New organizational structures
- Repositories, workflow, data archiving
- Finding relevant sources
- Self correcting databases
- Curation and funding
- Education and training

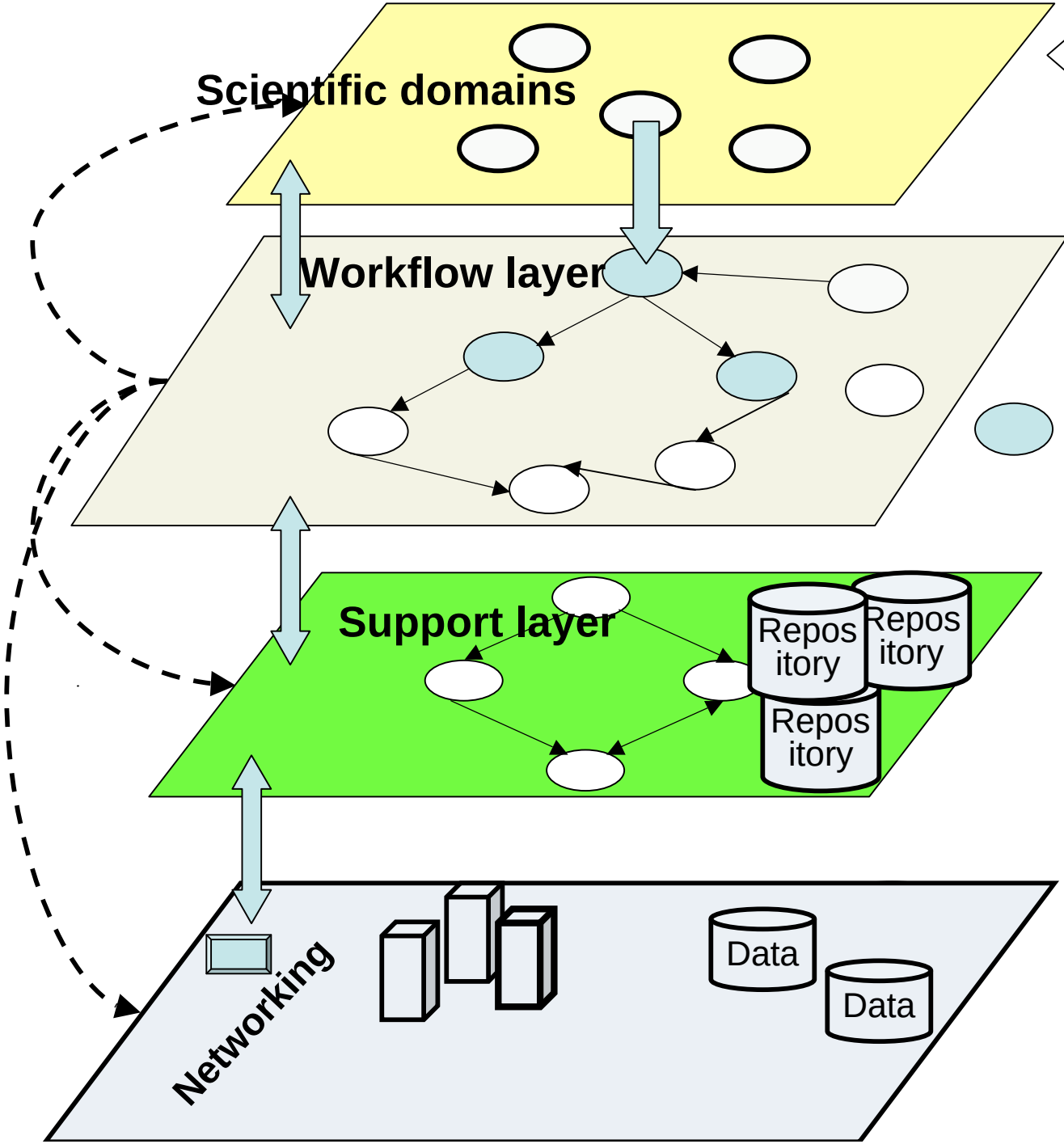
"*Without libraries, what have we? We have no past and no future*" -- Ray Bradbury

Critical opportunities for libraries in the next few years

1. **Data publishing** on the grid
2. **Data curation** - long term knowledge & culture preservation
   - Preserving knowledge - one of the most vital and yet rapidly changing functions
     - ensuring quality of information and archiving research data
3. **Machines -** the **next generation of readers**
   - ✓ Machines will "read" those new and old optimized book collections
4. Facilitating **global user workflows**
   - Collaboration is enlarging research boundaries and blurring disciplinary boundaries
5. **Enhancing the user experience** -- sense making and channel editing

A Digital Repository for the Future





Scientific Domains

(Generating Workflows)

Collaboratories

Applications & hybrid apps.

Toolsets, mashups,

(Workflows)

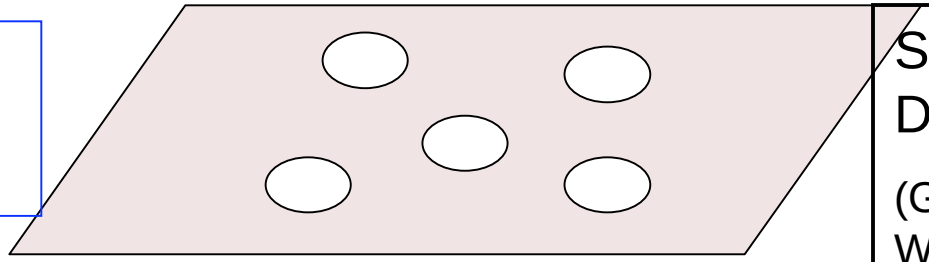
Digital library repositories

Data repositories

Data replication

Rick Luce, CIC, 5/12/08

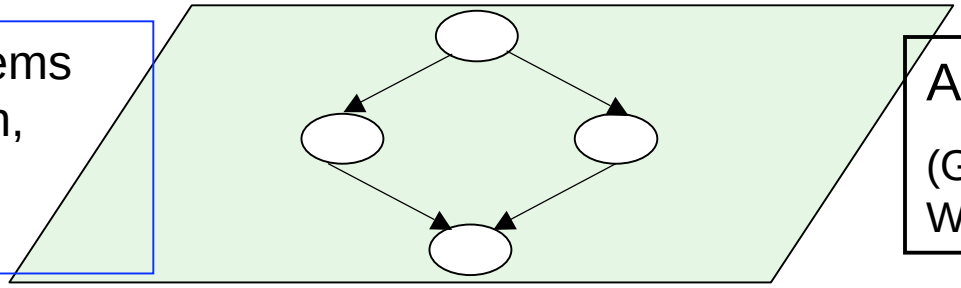
Project based science teams: interdisciplinary & trans-disciplinary



Scientific Domains

(Generating Workflows

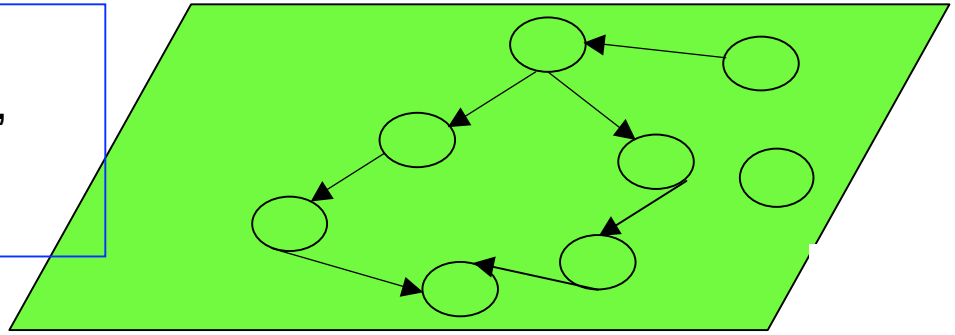
Domain experts, systems and application design, database experts, programmers,



Applications

(Generating Workflows

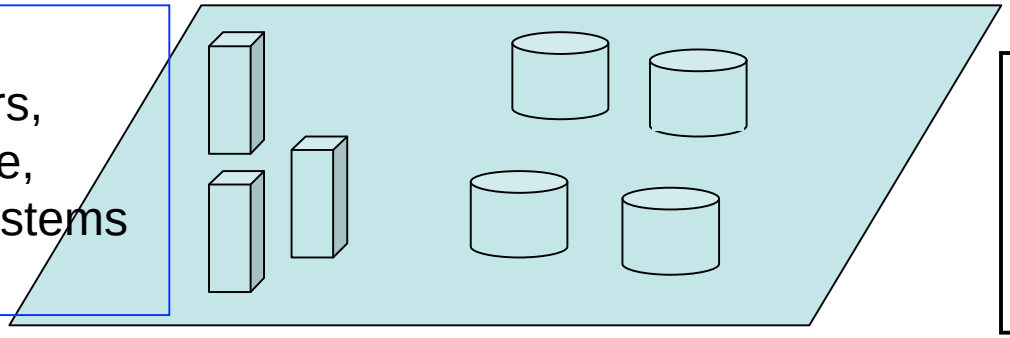
Digital library: information science, library science, domain experts, technologists



Repository layer

(Generating Workflows

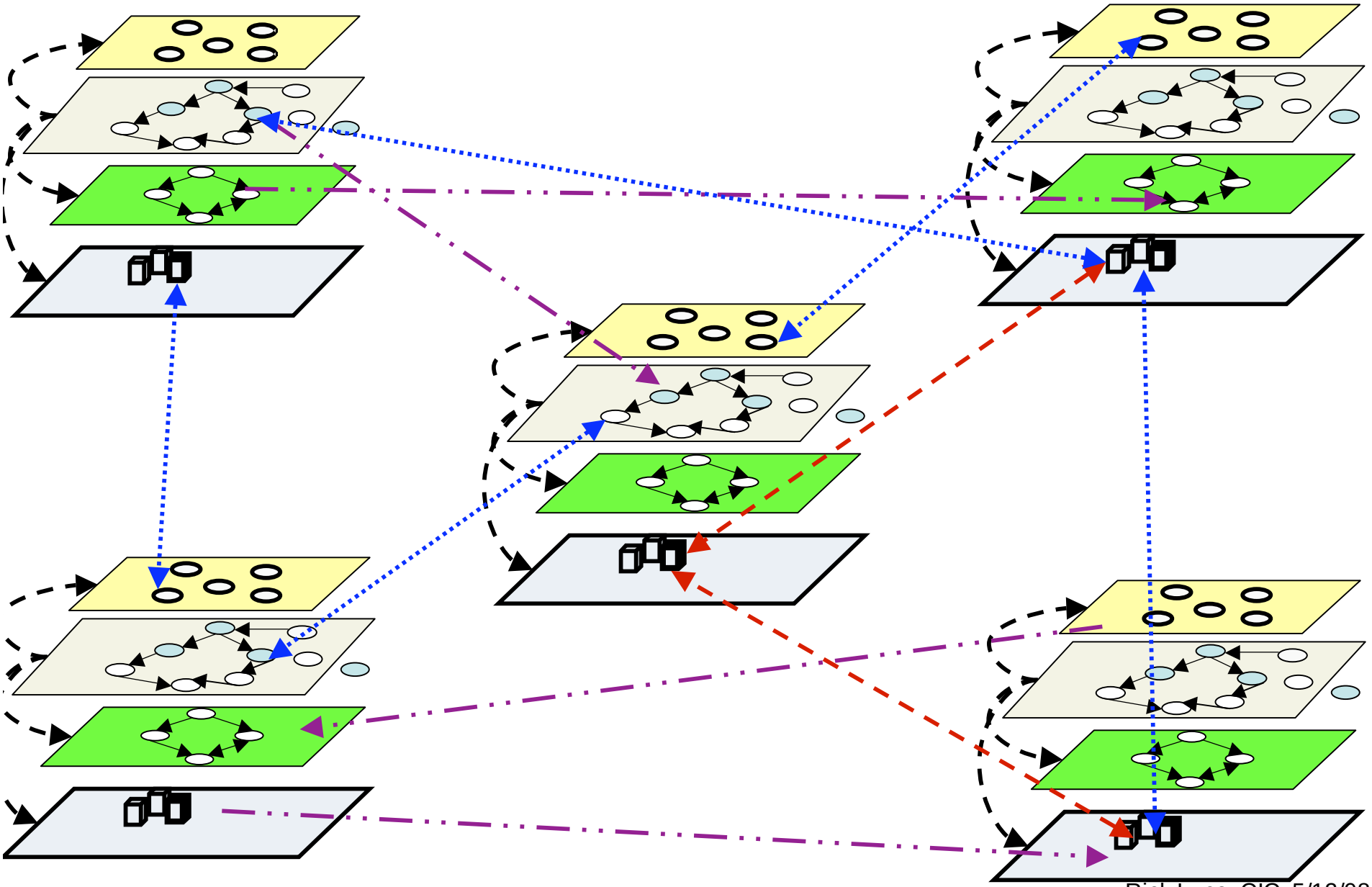
IT: network and storage engineers, computer science, technologists, systems experts



Data and infrastructure

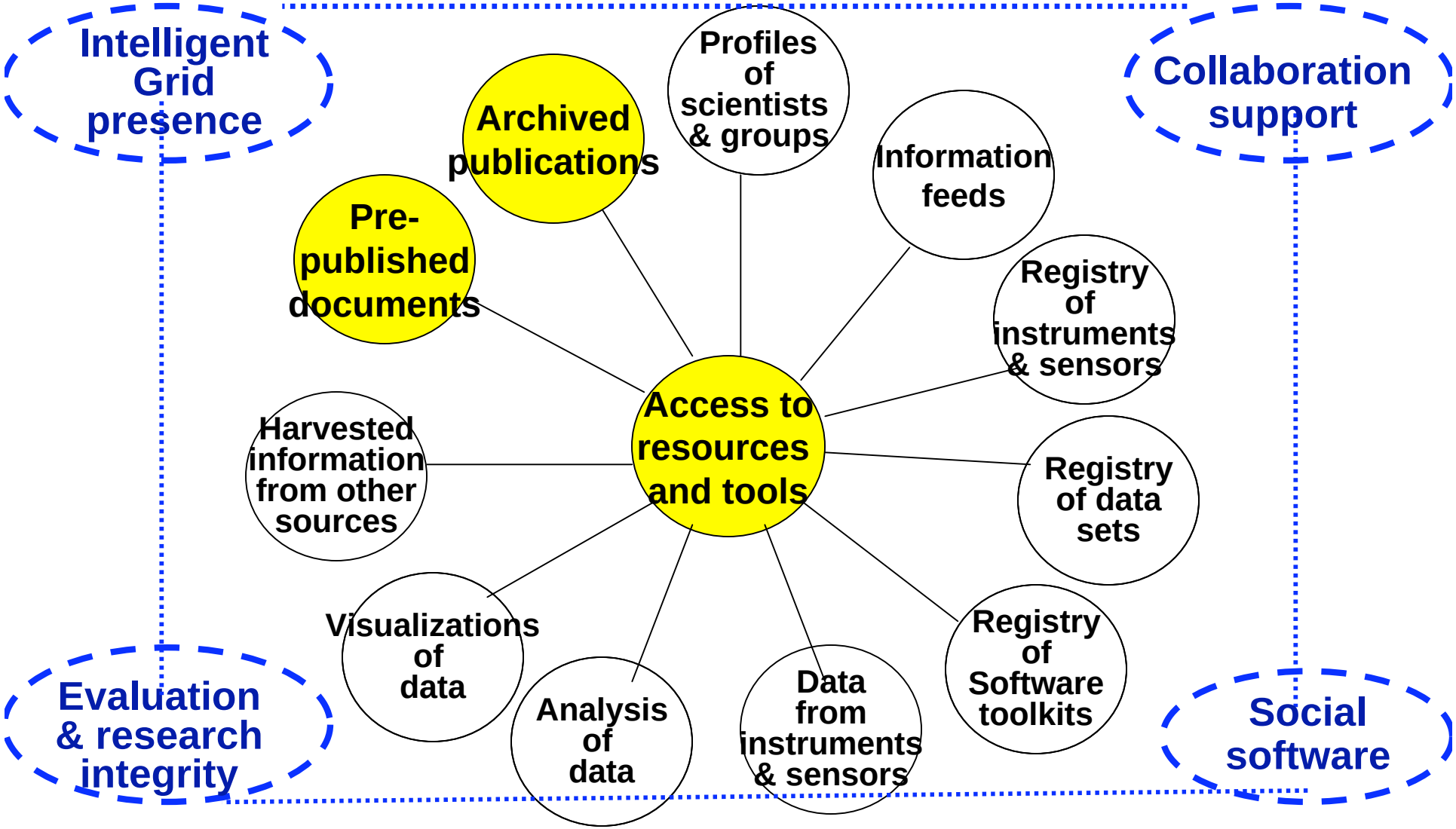
(Generating Workflows

Coordinating a on multi-institutional scale



Repositories, Workflow and Data Archiving

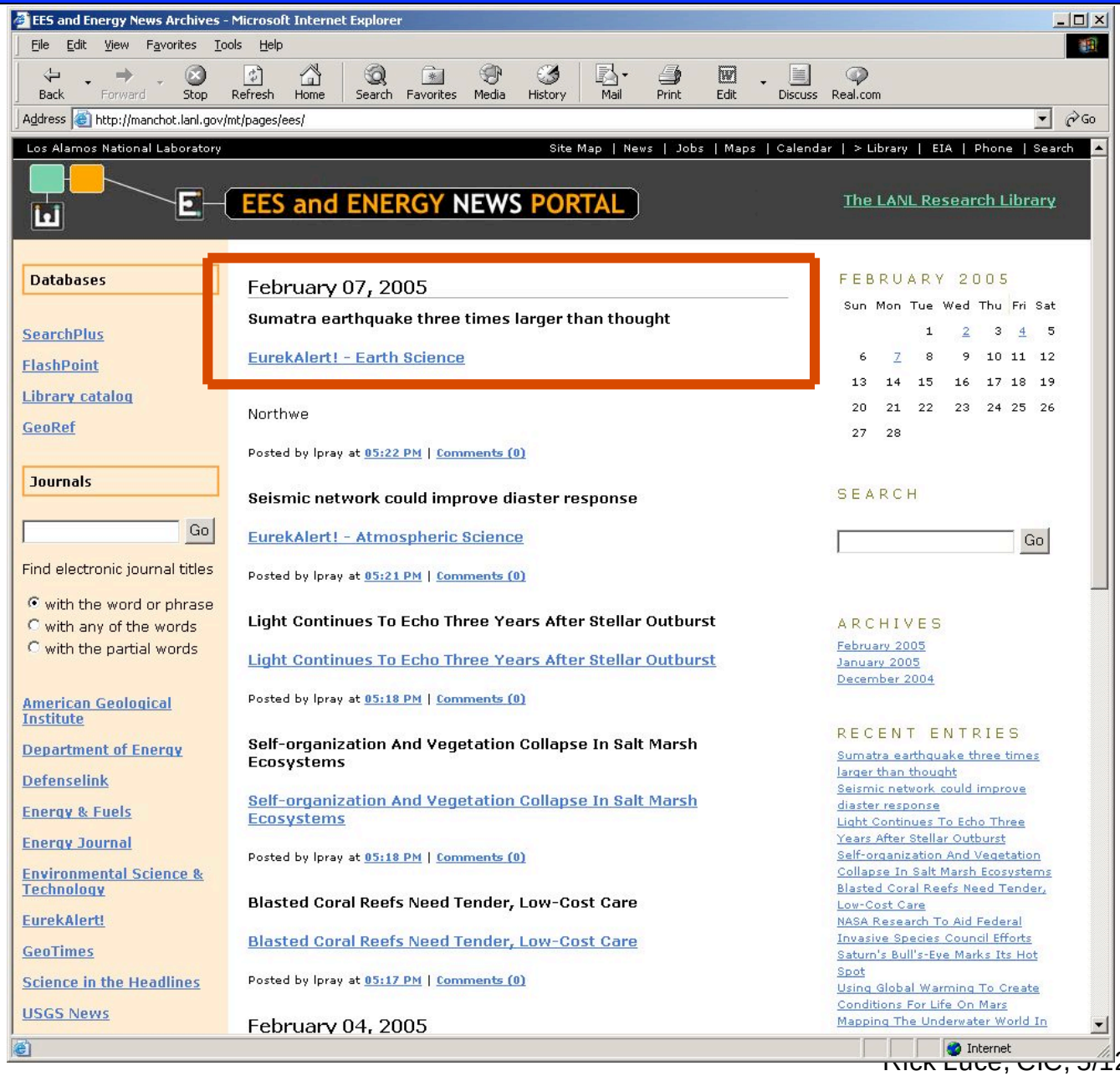
Enabling group-to-group interaction in persistent electronic spaces



Some Very Early Examples

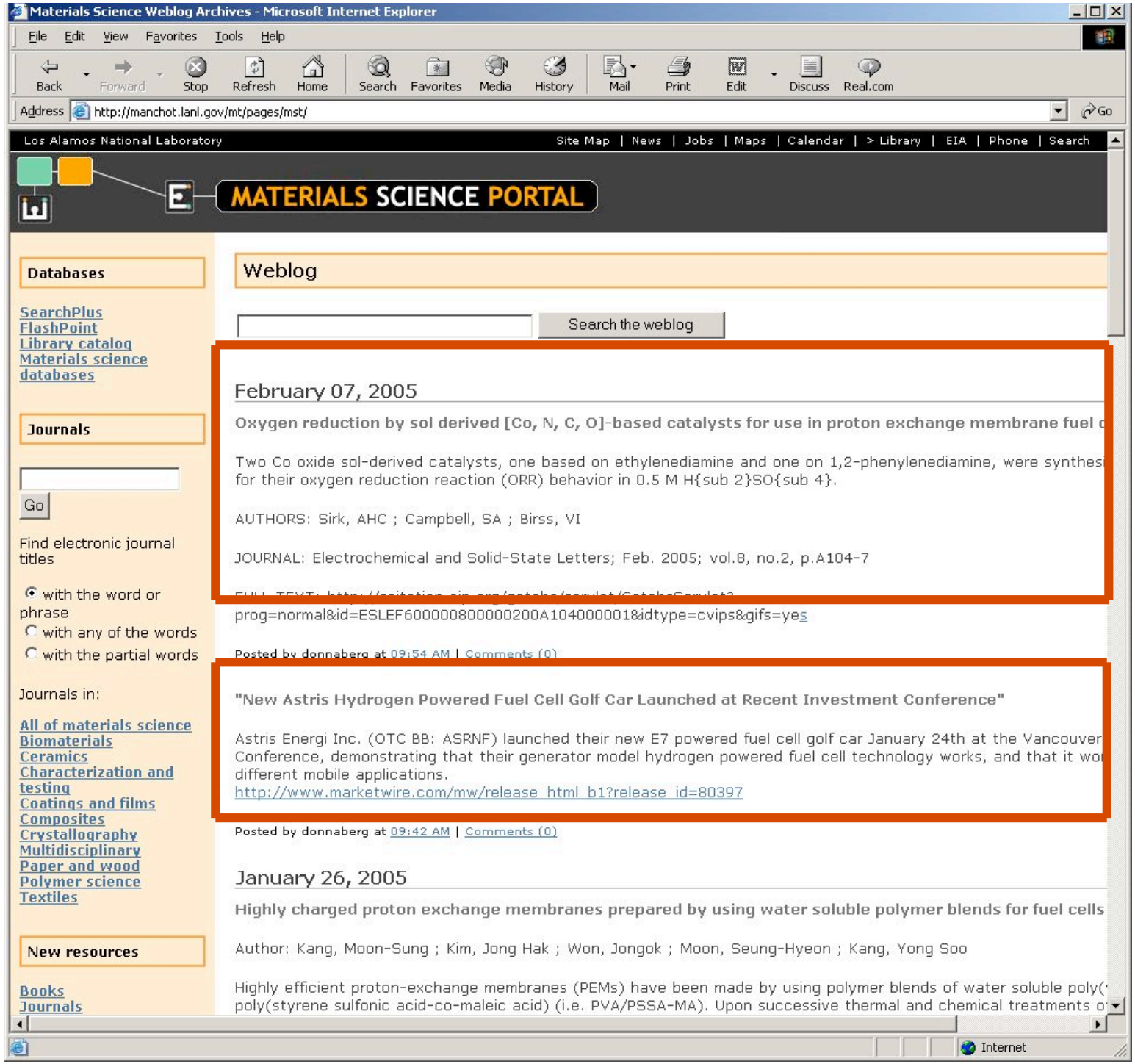
Weblog content:

EES and Energy portal includes news



Weblog content:

Materials Science portal: includes new articles and news



MATERIALS SCIENCE COLLAB

Databases

SearchPlus
FlashPoint
Library catalog
Materials science databases

Journals

Go

Find electronic journal titles

- with the word or phrase
- with any of the words
- with the partial words

Journals in:

All of materials science
Biomaterials
Ceramics
Characterization and testing
Coatings and films
Composites
Crystallography
Multidisciplinary
Paper and wood
Polymer science
Textiles

Crystallography journals, current tables of contents

Weblog

Post a new entry to the weblog

Search the weblog:

Search

APRIL 2004

Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

March 22, 2004

Example of a paper posted to the weblog

Title: Superconductivity in the non-oxide perovskite MgCNi3

Authors: He, T; Huang, Q; Ramirez, AP; Wang, Y; Regan, KA; Rogado, N; Hayward, MA; Haas, MK; Slusky, JS; Inumara, K; Zandbergen, HW; Ong, NP; Cava, RJ

Source: LA-UR 01-2223

Abstract: The interplay of magnetic interactions, dimensionality, and electronic correlations in producing superconductivity is one of the dominant themes in the current study of electronic ground states in complex materials. Originally thought to be detrimental for the formation of superconductors, magnetic interactions and

News RSS feeds

Industrial News Room - Materials & Material Processing

- Blackening Gel works on iron and steel substrate parts.
- Inclined Agitator eliminates dead spots.
- Processing System emulsifies, blends, and homogenizes.
- Aqueous Parts Washer cleans parts on assembly line.
- Multi-Reel Rewinder is suited for small floor spaces.
- Reclaim Extruder targets foam reprocessing applications.
- Adhesive Film conforms to curved areas of body.
- Fluid Bed Processors are offered on rental basis.
- Coating lets molders choose necessary level of release.
- Extrusion Die suits shear sensitive/low viscosity materials.
- Cast Film System

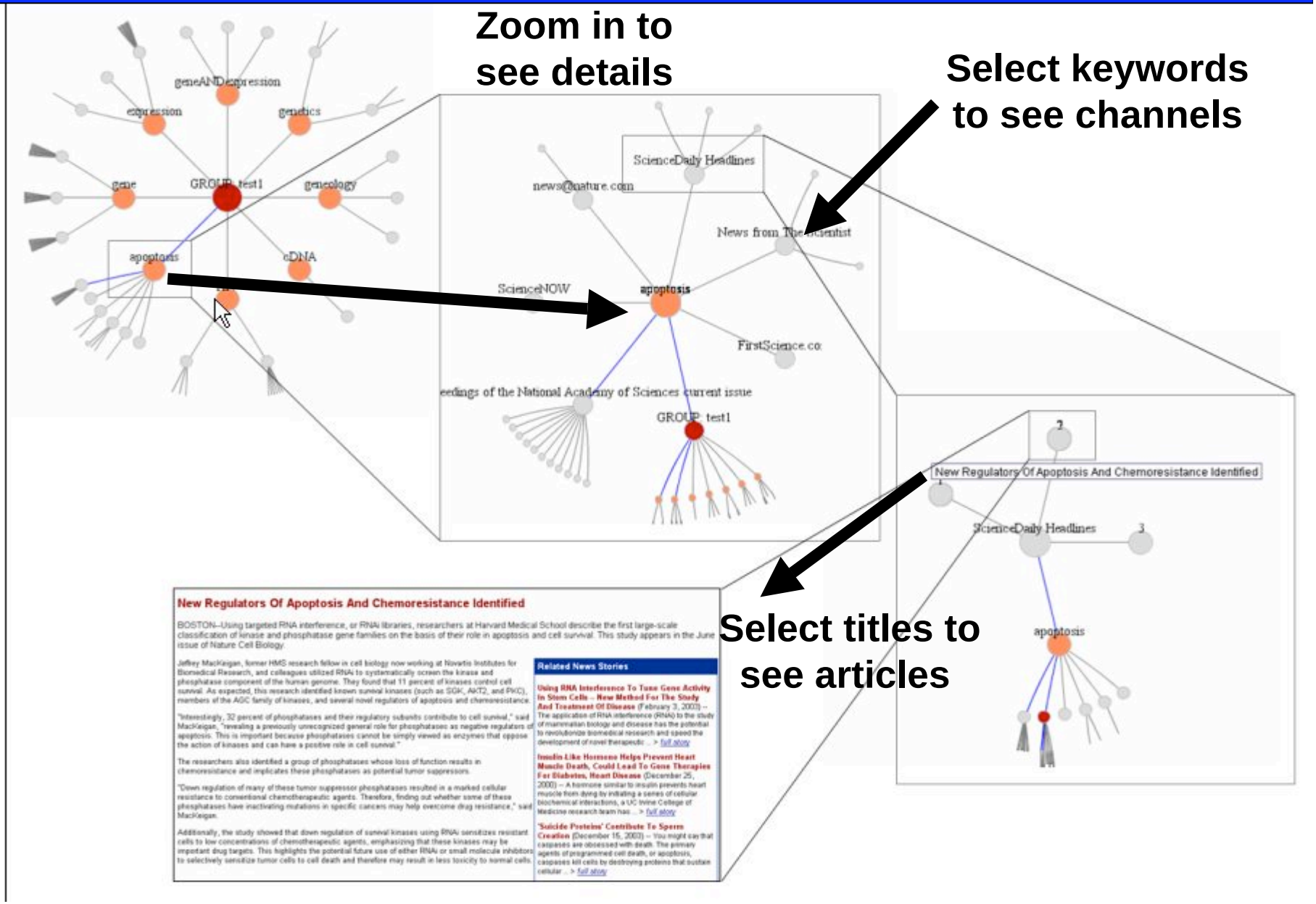
ScienceSifter @ LANL: creating a category feed

Profile a group's information needs

Create aggregated, filtered RSS feeds

- Aggregate source feeds
- Filter source feeds by keywords

ScienceSifter: exploring results



Rick Luce, CIC, 5/12/08

ScienceSifter: Viewing Feeds

DESCRIPTIVE FEEDS: Keyword related title and description are shown here

gene

Biology of the Cell

No new feeds are available

Cell

☐ And the Noise Played on: Stochastic Gene Expression and HIV-1 Infection -

Description: Blake, William J.; Collins, James J. Stochastic gene expression has been implicated in a variety of cellular processes, including cell differentiation and disease. In this issue of Cell, Weinberger et al. (2005) take an integrated computational-experimental approach to study the Tat transactivation feedback loop of HIV-1. They show that fluctuations in a key regulator, Tat, in an isogenic population of infected cells result in two distinct expression states corresponding to latent and productive HIV-1 infection. These findings demonstrate the importance of stochastic gene expression in molecular â€œdecision-making.â€RSS 2.0 feed provided by the LANL Research Library

☐ Stochastic Gene Expression in a Lentiviral Positive-Feedback Loop: HIV-1 Tat Fluctuations Drive Phenotypic Diversity -

Description: Weinberger, Leor S.; Burnett, John C.; Toettcher, Jared E.; Arkin, Adam P.; Schaffer, David V. SummaryHIV-1 Tat transactivation is vital for completion of the viral life cycle and has been implicated in determining proviral latency. We present an extensive experimental/computational study of an HIV-1 model vector (LTR-GFP-IRES-Tat) and show that stochastic fluctuations in Tat influence the viral latency decision. Low GFP/Tat expression was found to generate bifurcating phenotypes with clonal populations derived from single proviral integrations simultaneously exhibiting very high and near zero GFP expression. Although phenotypic bifurcation (PheB) was correlated with distinct genomic integration patterns, neither these patterns nor other extrinsic cellular factors (cell cycle/size, aneuploidy, chromatin silencing, etc.) explained PheB. Stochastic computational modeling successfully accounted for PheB and correctly predicted the dynamics of a Tat mutant that were subsequently confirmed by experiment. Thus, Tat stochastics appear sufficient to generate PheB (and potentially proviral latency), illustrating the importance of stochastic fluctuations in gene expression in a mammalian system.RSS 2.0 feed provided by the LANL Research Library

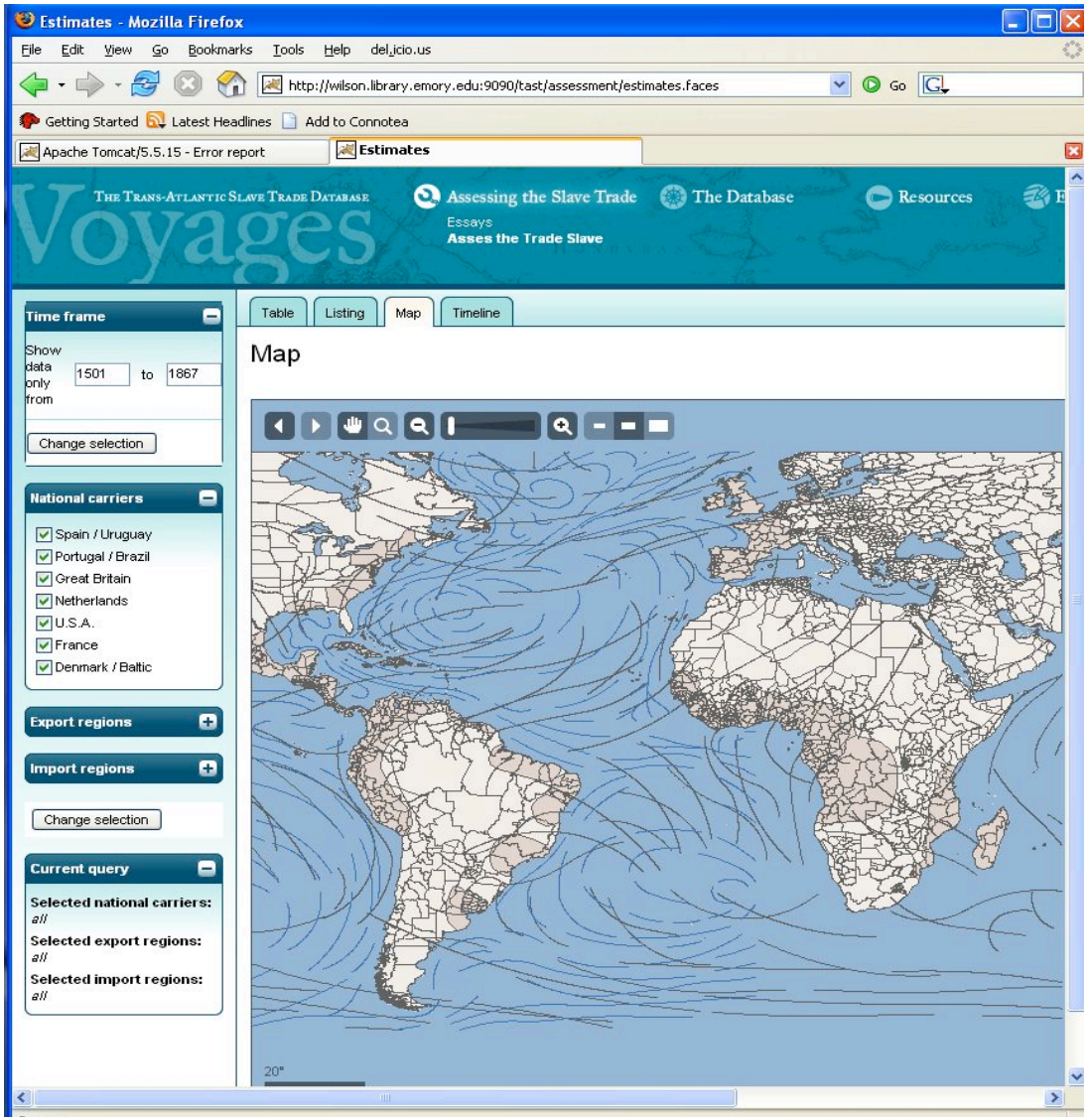
☐ Transcriptional Feedback of Neurospora Circadian Clock Gene by Phosphorylation-Dependent Inactivation of Its Transcription Factor -

Description: Schafmeier, Tobias; Haase, Andrea; KÃ¡ldi, Krisztina; Scholz, Johanna; Fuchs, Marc; Brunner, Michael SummaryThe circadian clock protein Frequency (FRQ) feedback-regulates its own expression by inhibiting its transcriptional activator, White Collar Complex (WCC). We present evidence that FRQ regulates the bulk of WCC through modulation of its phosphorylation status rather than via direct complex formation. In the absence of FRQ, WCC is hypophosphorylated and transcriptionally active, while WCC is hyperphosphorylated and transcriptionally inactive when FRQ is expressed. The

Voyages @ Emory (Transatlantic Slave Trade Database Online)

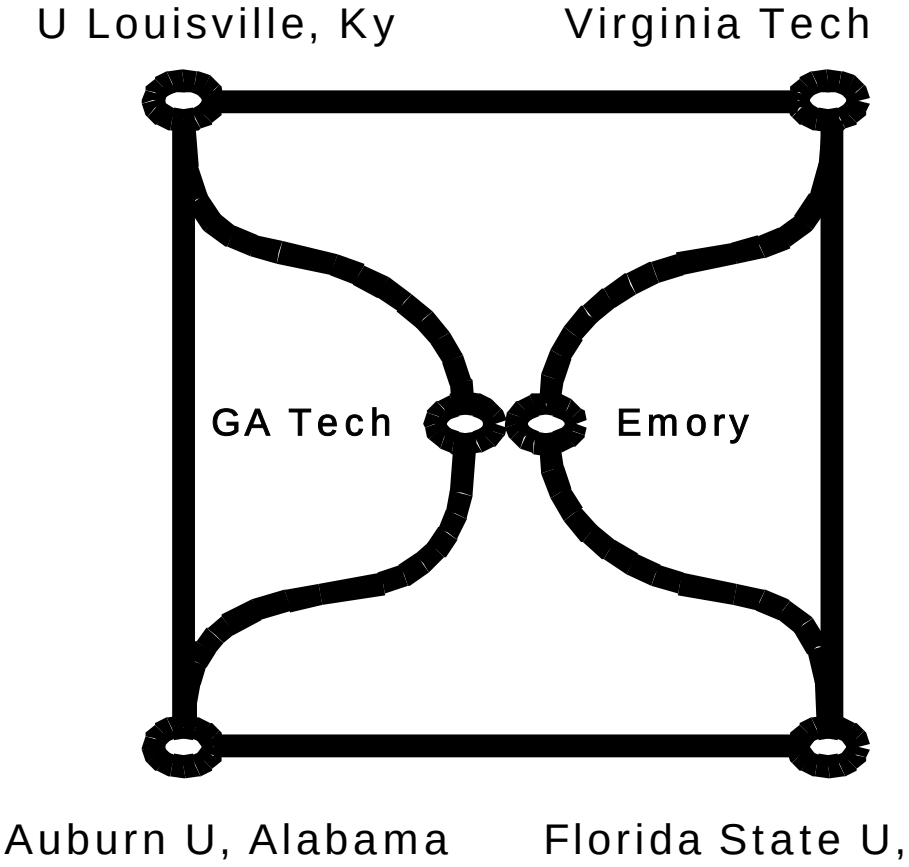
Integrated system for quantitative and geographic analysis for interdisciplinary inquiry

- Canonical database chronicling the five centuries of the transatlantic slave trade
- Statistical and geographic charting
- Digital culmination of decades of historical research by international team of scholars
- Creating web infrastructure for international collaboration to assemble primary research knowledge



Preservation: MetaArchive Project @ Emory

- Collaborative partnership: 6 universities, the National Digital Preservation Program, and the LOCKSS Alliance
- Secure, distributed network for digital archives with:
  - conspectus building
  - organizational agreements
  - content selection
  - format migration
  - technical training
- Producing new open source applications of the LOCKSS software for distributed institutional archives



Early Direction at Emory

- Grey space synergy between IT / Library
- Library = laboratory for the 2015 classroom
- Collective problem solving (vis-à-vis open source)
- Embed eResearch staff with projects, customers
- Break down silos - leverage resources across organizational boundaries

Revolution happens when it cannot be contained by status quo institutions

Partnerships: Outside In & Inside Out

Libraries are where our social networks and technological networks overlap

Turning libraries outside in and staffing inside out

- Bringing classrooms inside, integrating learning spaces
- Faculty partnerships inside: laboratories
- Moving staff outside → embedded where users live

Partners: industry, faculty, libraries, …everywhere

Luce, R. *A New Value Equation Challenge: The Emergence of eResearch and Roles for Research Libraries*. CLIR, 2008.

Bottom Line

In this arena libraries are a microcosm of the university

- *Requires new organizational structures* –

… at all levels in the academy

…with a different culture

… and new hybrid distributed organizational structures

…geared to innovation and experimentation

Historical Perspective of Library Cooperatives

Library collaboration: resource sharing networks, multi-types driven by:

1. Duplicative work (e.g., OCLC); or
2. Geography (local & regional networks); or
3. Common shared characteristics

Next phase shift requires an expanded mission of shared purpose

Requires collaboration and collective action

Taking Collaboration to the Next Level

Challenges where we come up short:

- Scale, speed, agility, resources, …
- *Cloud computing* and shared distributed resources involves linking together large pools of systems to provide IT services

- Collaboration networks based on **'cloud' principles**

Cloud collaborating for libraries?

Different Dimensions for Collaboration

Working out collaborative environments:

- Behavior norms, protocols of project engagement
- Intellectual property
- Workflow and handoffs
- Financial investments
- People and staffing
- Project management & leadership
- Sharing distributed infrastructure

In a knowledge economy, characterized both by collaboration and competition in science, the enabling infrastructure and support systems are fundamental to competitiveness and scientific leadership.

Thinking outside the box

Funding vs. Disruption

Could it be that we are well enough funded to be comfortable with our traditional roles?

- Think about disruption and discontinuous change

  India - 2,000 new universities & colleges in 5 years

Can We Think Innovatively?

Requires shifting from focus on products (e.g., reference) to process (e.g., team science)

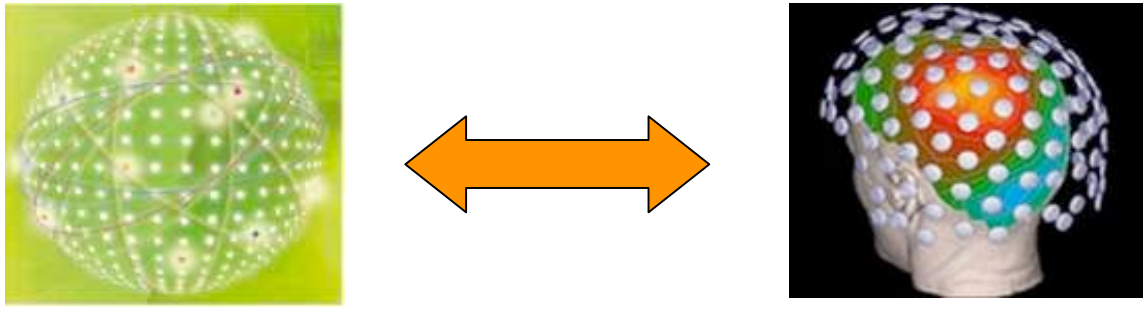
Supporting the knowledge workplace

- **Shared ideas**: discussion DB's, intranets, portals …
- **Shared presence**: social software, videoconferencing, networked virtual worlds, avatars
- **Shared creation**: workflow, CMS, web conferencing, collaborative design

Digital Libraries would become true information utilities

The future belongs not to those that merely navigate us through cyberspace, nor those who populate it with data. Rather it belongs to those who help us make sense of all that is available to us.

Concluding thoughts … on the road ahead



We must do much more than aggregate and provide access to digital scientific information … Our job now is to wire people's brains together so that sharing, reasoning, and collaboration become part of everyday work.

Rick Luce, *Bits*, April 1998