



Disciplinary Differences in e-Research: An Information Perspective

Christine L. Borgman

Professor & Presidential Chair in Information Studies

University of California, Los Angeles

and

Academic Visitor, Oxford Internet Institute



One cyberinfrastructure for all?

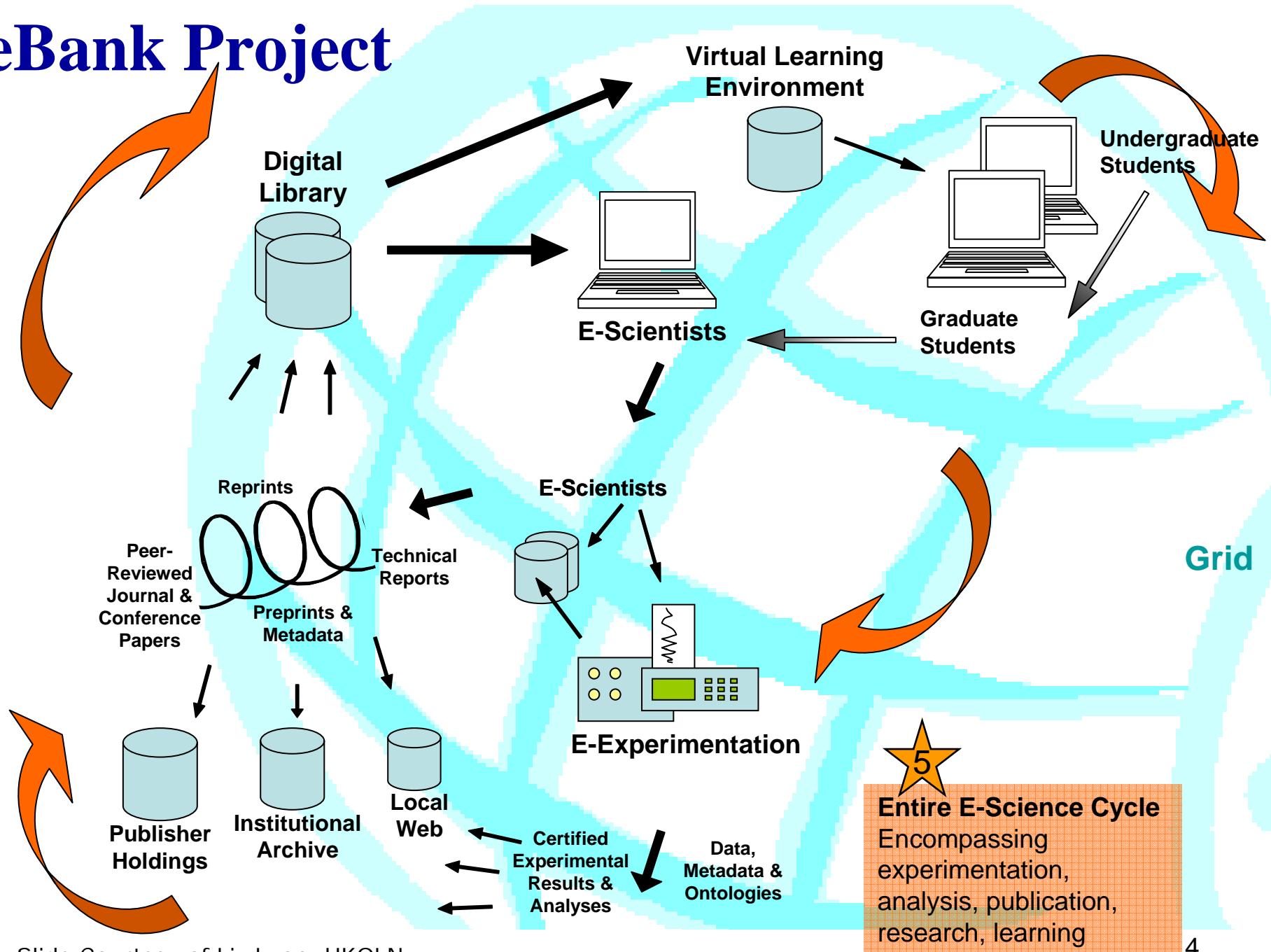
- ✧ e-Science
- ✧ e-Social Science
- ✧ e-Humanities
- ✧ e-research
- ✧ e-learning
- ✧ e-geosciences
- ✧ e-medicine
- ✧ e-engineering...



e-Research and Cyberinfrastructure Goals

- ✧ to enable research and learning that is
 - ✧ data-intensive
 - ✧ information-intensive
 - ✧ distributed
 - ✧ collaborative
 - ✧ multi-disciplinary
- ✧ to create a “value chain” of data and documents
 - ✧ to facilitate replication and verification of research
 - ✧ to enable new models, combinations of information
 - ✧ to manage the “data deluge”

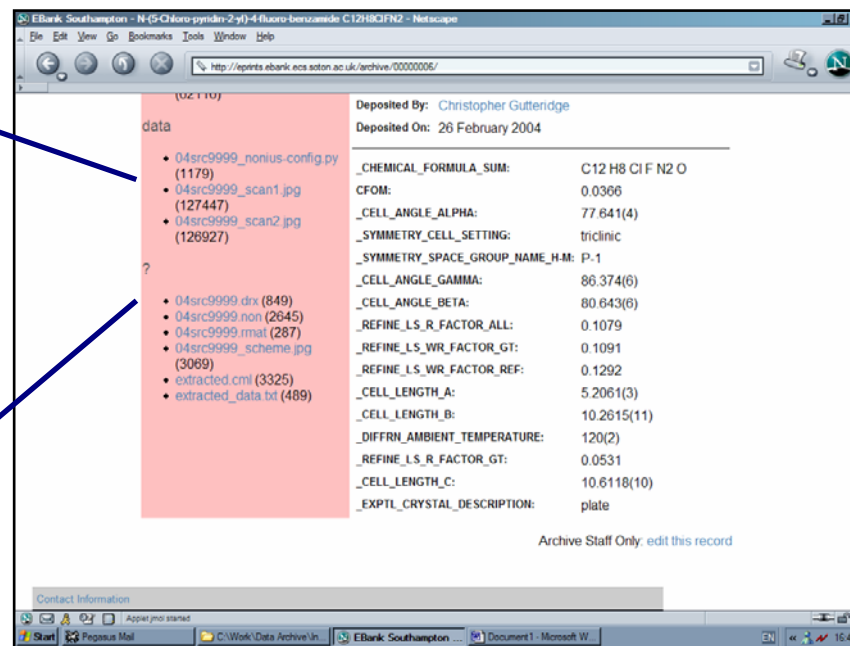
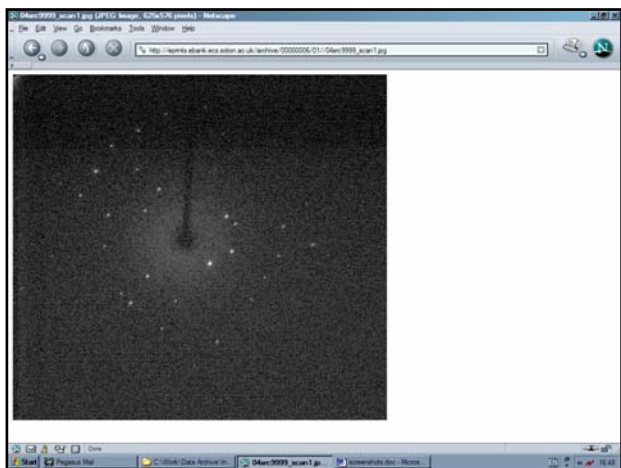
eBank Project



Slide Courtesy of Liz Lyon, UKOLN

Crystallographic e-Prints

➤ Direct Access to Raw Data from scientific papers



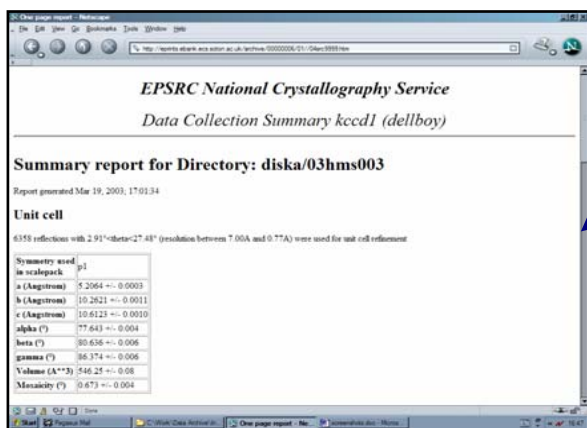
data

- 04src9999_nonius-config.py (1179)
- 04src9999_scan1.jpg (127447)
- 04src9999_scan2.jpg (126927)

Deposited By: Christopher Guttenidge
Deposited On: 26 February 2004

_CHEMICAL_FORMULA_SUM:	C12 H8 Cl F N2 O
CFOM:	0.0366
_CELL_ANGLE_ALPHA:	77.641(4)
_SYMMETRY_CELL_SETTING:	triclinic
_SYMMETRY_SPACE_GROUP_NAME_H_M:	P-1
_CELL_ANGLE_GAMMA:	86.374(6)
_CELL_ANGLE_BETA:	80.643(6)
_REFINE_LS_R_FACTOR_ALL:	0.1079
_REFINE_LS_WR_FACTOR_GT:	0.1091
_REFINE_LS_WR_FACTOR_REF:	0.1292
_CELL_LENGTH_A:	5.2061(3)
_CELL_LENGTH_B:	10.2615(11)
_DIFFRN_AMBIENT_TEMPERATURE:	120(2)
_REFINE_LS_R_FACTOR_GT:	0.0531
_CELL_LENGTH_C:	10.6118(10)
_EXPTL_CRYSTAL_DESCRIPTION:	plate

Archive Staff Only [edit this record](#)



EPSRC National Crystallography Service
Data Collection Summary *kccd1 (dellboy)*

Summary report for Directory: **diska/03hms003**
Report generated Mar 19, 2003, 17:01:34

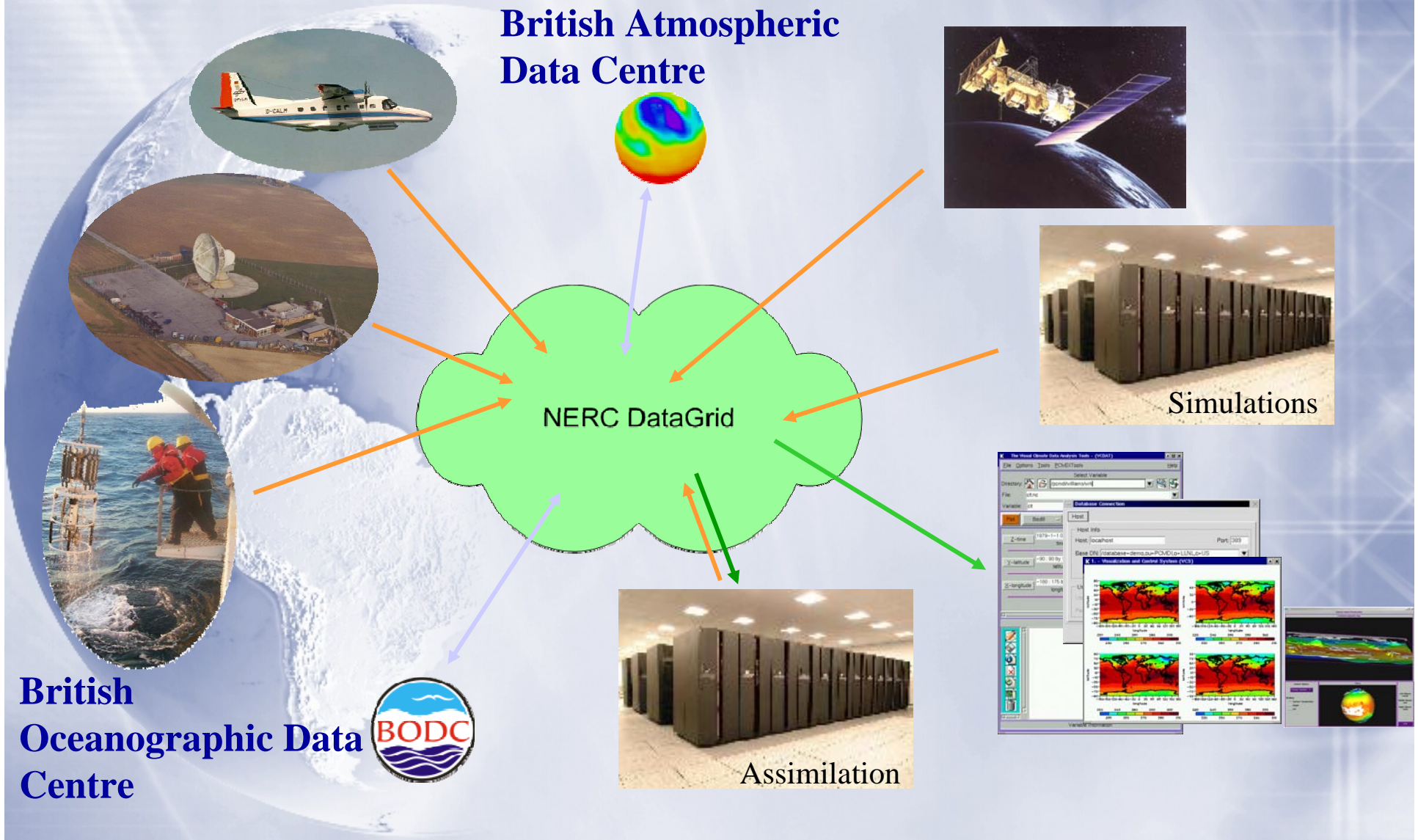
Unit cell

6158 reflections with 2.91° data (27.48° resolution between 7.00Å and 0.77Å) were used for unit cell refinement

Symmetry used:	p1
in scalepack:	p1
a (Angstrom):	5.2064 ± 0.0003
b (Angstrom):	10.2621 ± 0.0011
c (Angstrom):	10.6123 ± 0.0010
alpha (°):	77.641 ± 0.001
beta (°):	80.636 ± 0.006
gamma (°):	86.374 ± 0.006
Volume (Å ³):	546.25 ± 0.08
Minisity (°):	0.673 ± 0.004

Raw data sets can be very large and these are stored at National Datastore using SRB server

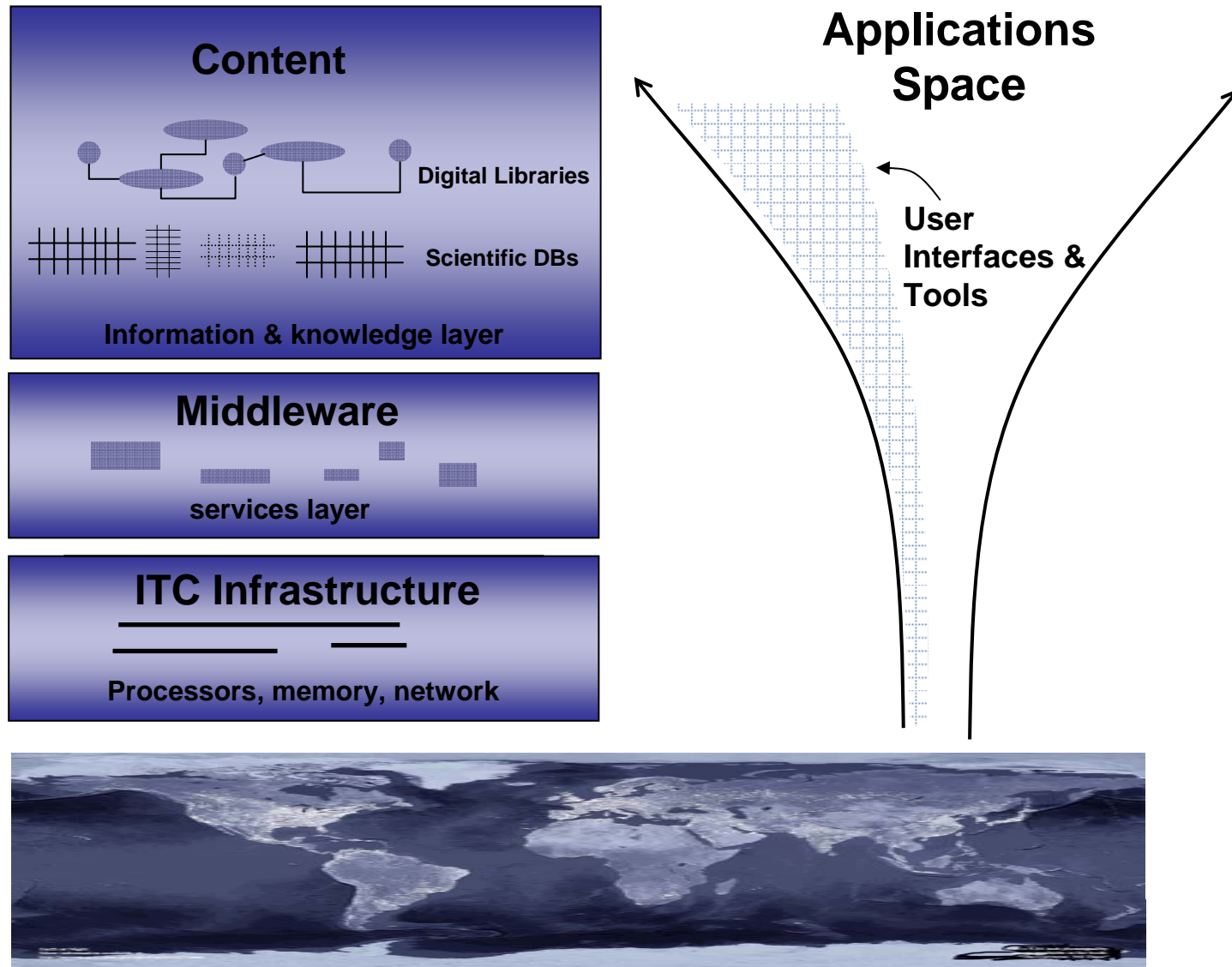
Complexity + Volume + Remote Access = Grid Challenge



**British
Oceanographic Data
Centre**



Cyberinfrastructure: Layered Model



Slide courtesy of Norman Wiseman, JISC



An infrastructure OF or FOR information?

✧ OF information:

- ✧ building a framework to support any kind of information
- ✧ meaning of information is in the eyes of the sender and receiver

✧ FOR information:

- ✧ building a framework to provide context for interpretation, use, re-use
- ✧ representations of information are abstractions, not literal pictures of objects in the “real world”
- ✧ meaning of information is negotiated and political



Contents of an infrastructure FOR information

✧ Documents

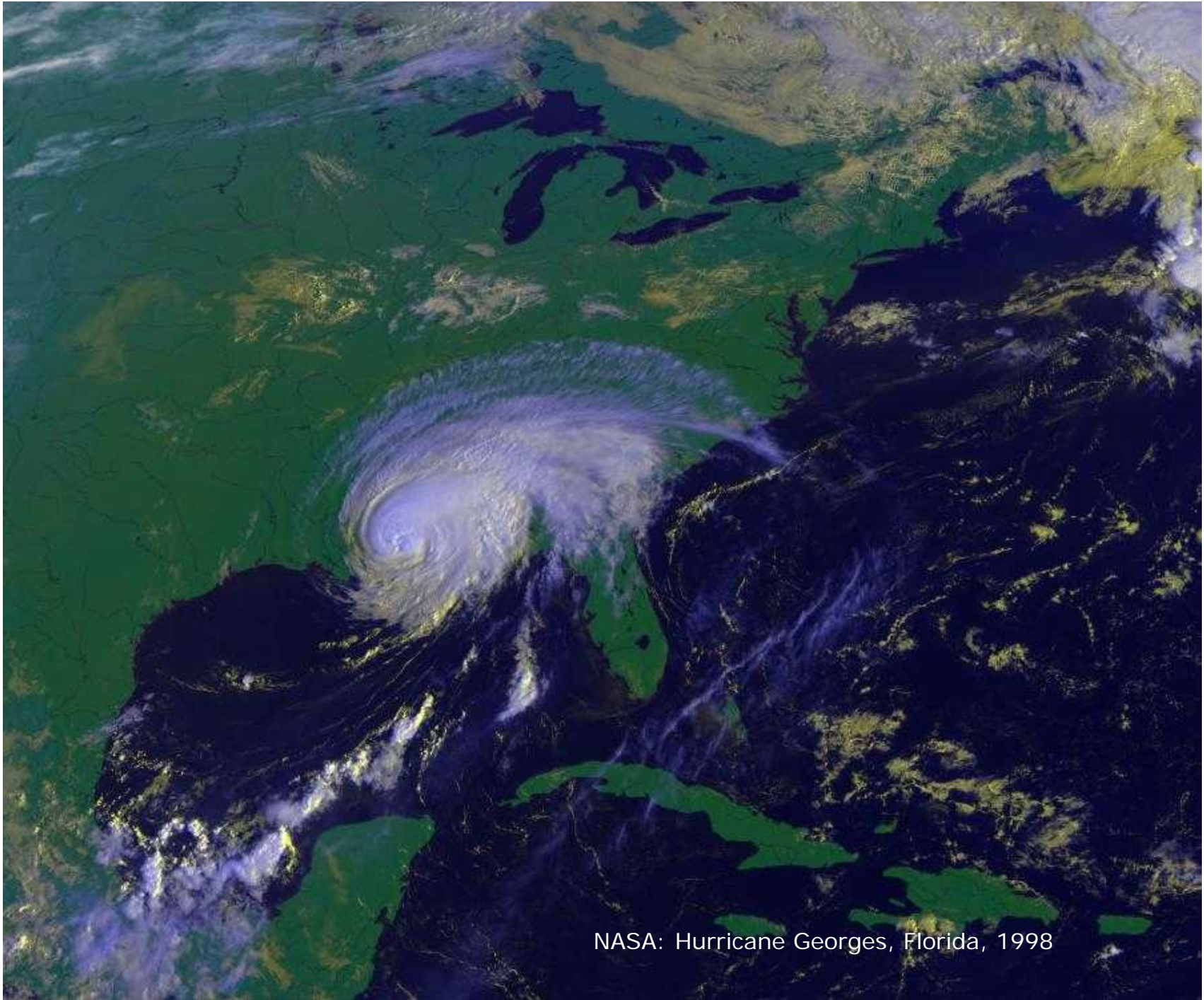
- ✧ Scholarly publications, theses, dissertations
- ✧ Pre-prints, reprints, reports, working papers
- ✧ General publications: government, mass media, fiction, non-fiction

✧ Data

- ✧ Experiments, observations, instrument output, geospatial coordinates
- ✧ Interviews, surveys, census, cultural artifacts
- ✧ Records of individuals, governments, organizations

✧ Composite objects

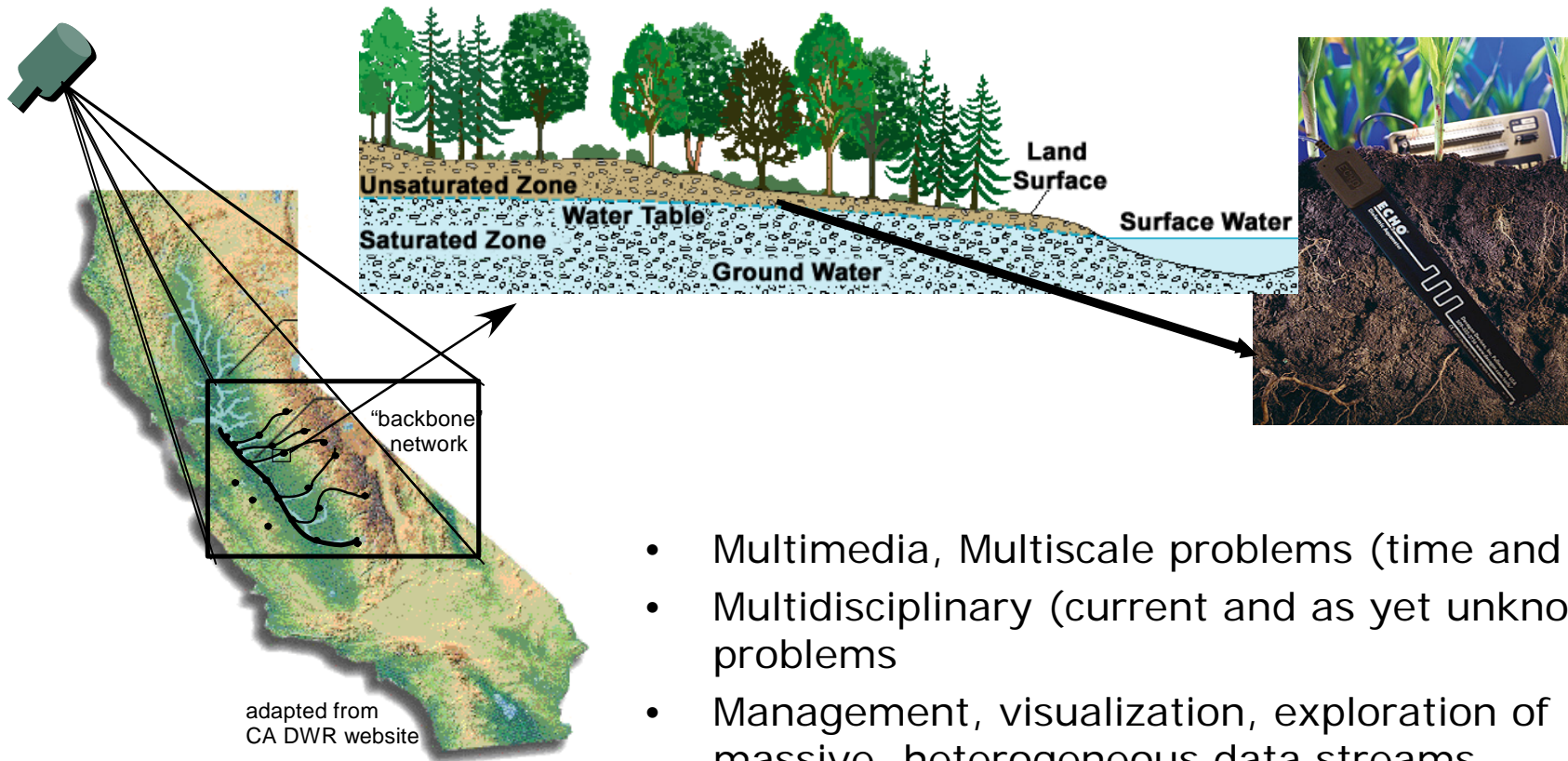
- ✧ Studies of habitat: environmental data, observations, field notes, biological data on individual species of flora and fauna
- ✧ Studies of political behavior: surveys, media reports, voting records, analytical reports, speeches
- ✧ Virtual reality models of archaeological sites: visualizations, geospatial coordinates, scholarly documentation



NASA: Hurricane Georges, Florida, 1998



Data models for habitat monitoring and sensor networks



- Multimedia, Multiscale problems (time and space)
- Multidisciplinary (current and as yet unknown) problems
- Management, visualization, exploration of massive, heterogeneous data streams



http://www.hpolar.org/

goldman s...

Apple Amazon Yahoo! News MyAthens UCLA Librar...e Databases OED online .Mac eBay

History and Politics Out Lo...

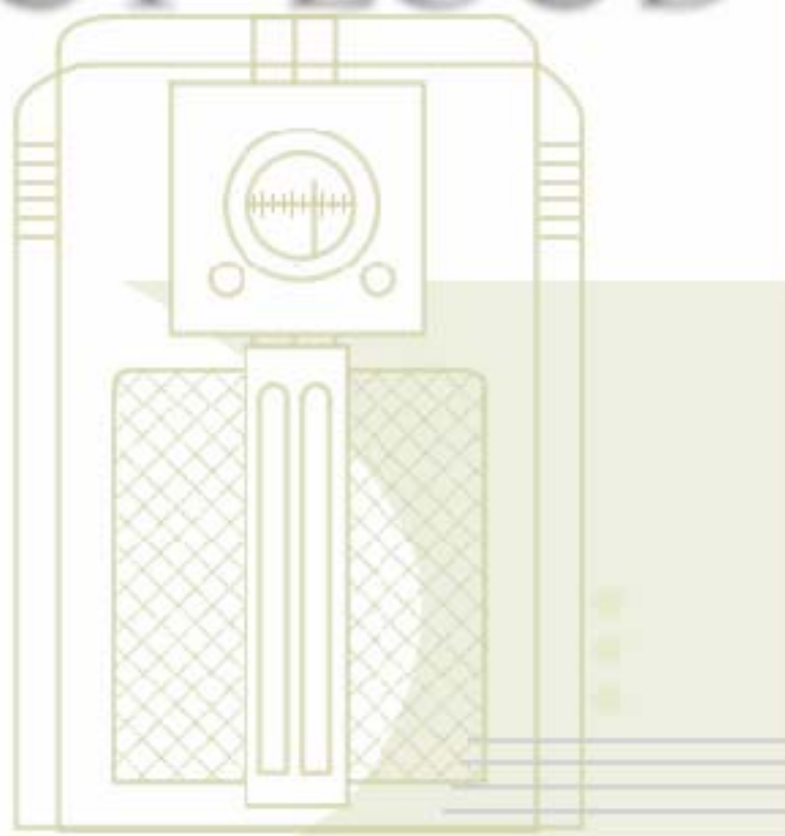
Now available:



Speeches of Rev. Martin Luther King, Jr. and speeches from the 1963 Civil Rights March in Washington D.C.

Selected Nixon Watergate recordings and transcripts now available. "Cancer on the Presidency" and "Smoking Gun" recordings and more...

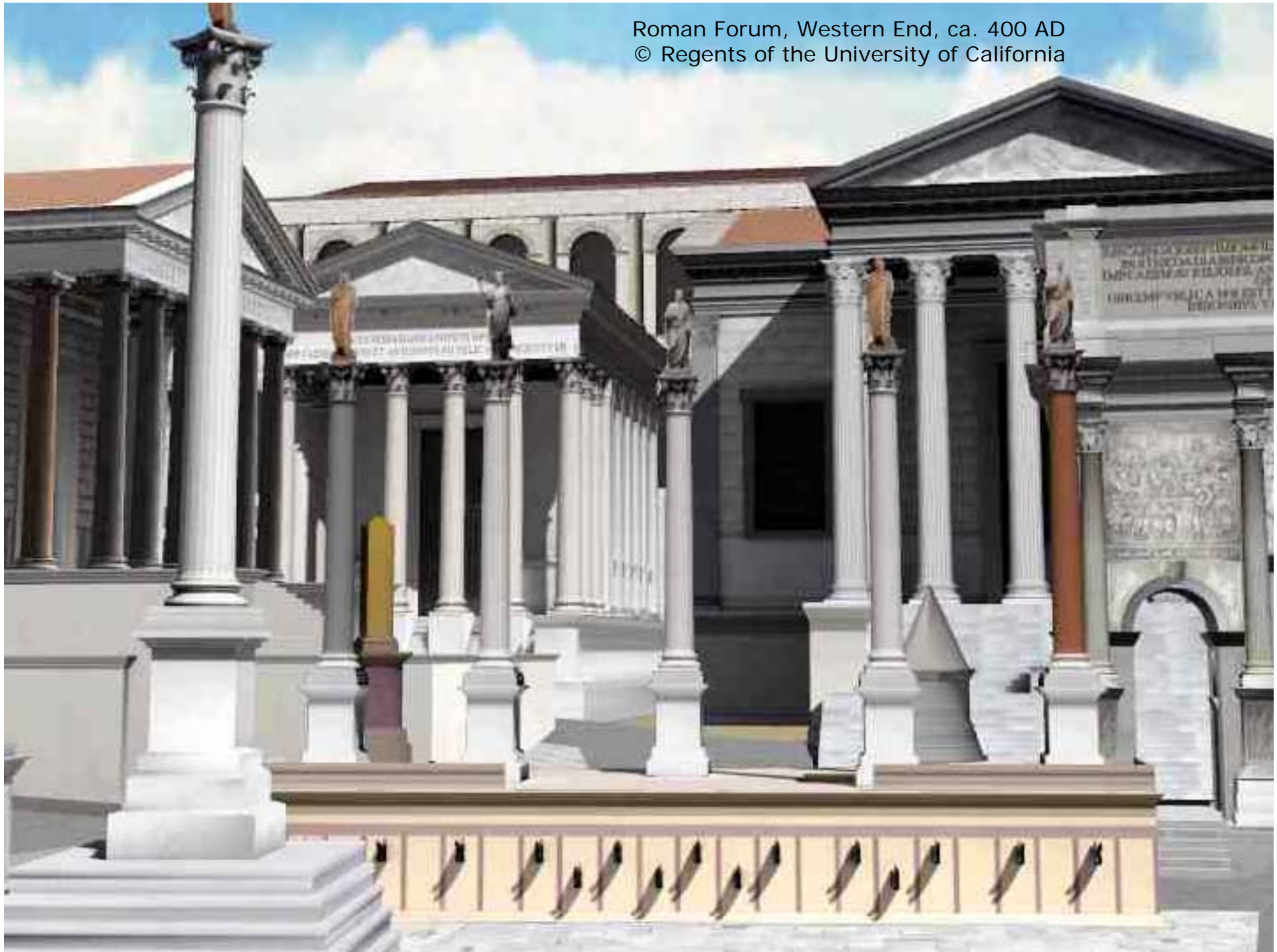
history and politics **OUT LOUD**



History and Politics Out Loud (HPOL) is a searchable archive of politically significant audio materials for scholars, teachers and students. HPOL is a component of "Historical Voices" funded by the National Endowment for the Humanities in partnership with Michigan State University.

www.hpolar.org

Roman Forum, Western End, ca. 400 AD
© Regents of the University of California





Some research questions

- ✧ Whose value chains are best served by an information infrastructure?
- ✧ Which information should be captured and curated?
- ✧ Which information is most likely to be shared or re-used?
- ✧ How does information creation, use, and sharing vary within and between fields?
- ✧ How can context be provided to facilitate use, interpretation, sharing of information?



Documents

- ✧ Common across disciplines
 - ✧ Scholarly publications as input to research
 - ✧ Peer review as quality control mechanism
 - ✧ Replication or verification of empirical research
 - ✧ Curation of scholarly publications necessary for long term access
 - ✧ Access via bibliographic records (metadata)
 - ✧ Libraries responsible for curation of print scholarly publications
- ✧ Linking adds context
 - ✧ Documents and data
 - ✧ Links between documents, e.g., citations



Documents

- ✧ Differences within and between disciplines
 - ✧ Collaboration and co-authorship
 - ✧ Highest in sciences, lowest in humanities
 - ✧ Related to funding, use of information technologies
 - ✧ Time frame for usefulness
 - ✧ Sciences
 - ✧ Shorter half-life of documents, citation-decay curve
 - ✧ Publish journal articles, reports, conference papers
 - ✧ Most of requisite documents available online
 - ✧ Humanities
 - ✧ Longer half-life of documents, citation-decay curve
 - ✧ Publish books, journal articles, conference papers
 - ✧ Small portion of requisite documents available online



Usefulness of documents and composite objects

✧ Sciences

- ✧ Reliance on current resources
- ✧ Value of documents, publications decays quickly
- ✧ Large audience, short period of time

✧ Humanities

- ✧ Reliance on current and historical resources
- ✧ Value of documents, composite objects decay slowly
- ✧ Large audience, long period of time



Data

- ❖ Common across disciplines

- ❖ Generating large volumes of data

- ❖ Science: sensors, satellites, instruments

- ❖ Social science: surveys, observations

- ❖ Humanities: digital libraries of cultural data

- ❖ Using their own data and that of others

- ❖ Scientists mining vast datasets of observations

- ❖ Social scientists combining statistical, textual, modeling data

- ❖ Humanists combining current and historic cultural data



Data

- ✧ Common across disciplines
 - ✧ Sharing
 - ✧ Collaborative, multi-disciplinary, multi-institutional projects
 - ✧ Common metadata standards *within* fields
 - ✧ Tools
 - ✧ Data visualization, pattern seeking
 - ✧ Capture, curation
 - ✧ Intellectual property
 - ✧ Agreements on ownership are complex
 - ✧ Complexity increases with number of partners, types of data
 - ✧ Quality control standards and practices needed
 - ✧ Few equivalents to peer review of publications
 - ✧ “Data that get used, improve”



Data

- ✧ Differences within and between disciplines
 - ✧ Number and variety of data types
 - ✧ Crystallography: molecules, protein structures
 - ✧ Political science: opinion polls, surveys, voting records, media reports, census data, local government records
 - ✧ Art history: letters, archival records, x-rays of paintings, documentary histories, chemical samples
 - ✧ Ability to identify all potential data sources
 - ✧ Sciences (e.g., atoms, molecules, genomes): high
 - ✧ Social sciences (e.g., large social surveys): medium
 - ✧ Humanities (e.g., cultural records): low
 - ✧ Agreement on representation
 - ✧ Chemistry: high agreement on representations of molecules
 - ✧ Social surveys: variable names specific to each study
 - ✧ Art history: low agreement on artists' names



Data

- ✧ Differences within and between disciplines
 - ✧ Sensitivity of data: privacy, confidentiality
 - ✧ Sciences (e.g., atoms, molecules, genomes): low
 - ✧ Humanities (e.g., cultural records): medium
 - ✧ Social sciences (e.g., interviews, observations): high
 - ✧ Medicine (e.g., patient records): high
 - ✧ Economic (resale) value of data
 - ✧ Chemistry: very high
 - ✧ Stock market, geospatial: time dependent
 - ✧ General social surveys: low
 - ✧ Cultural artifacts (e.g., images): high



Data

- ✧ Differences within and between disciplines
 - ✧ Obligations to deposit data
 - ✧ Some sciences, health: Deposit mandated by funding agency or journal
 - ✧ Some social sciences fields: Deposit mandated or encouraged
 - ✧ Humanities: Individual control is norm
 - ✧ Most fields: Journals will not accept supplemental data
 - ✧ Agreement on data repositories
 - ✧ Field- and discipline specific
 - ✧ Held by individual investigators
 - ✧ No repository may exist



Context and interpretation

- ✧ Scholarly publications *provide* context
 - ✧ Literature review, history of problem, definitions of terms
 - ✧ Theory, hypotheses, goals
 - ✧ Research method, discussion of results
- ✧ Repositories *remove* context
 - ✧ Data elements
 - ✧ Variable names
 - ✧ Instrument readings
 - ✧ Numerical data
 - ✧ Images, descriptions of artifacts



Sharing data

✧ Incentives to share data

- ✧ Tradition of “open science” to promote access, transparency
- ✧ Ability to replicate, compare results
- ✧ Ability to ask new questions by mining datasets
- ✧ Establish trust and reciprocity within a research group
- ✧ Requirement of some funding agencies, journals

✧ Incentives *not* to share data

- ✧ Rewards for publication, not for data management
- ✧ Benefits of contributing data may accrue to other parties
- ✧ Risks of misinterpretation of your data
- ✧ Risks of losing control over data
- ✧ Risks of loss of intellectual property



Summary-1

- ✧ e-Research goals: value chain of information-intensive, multi-disciplinary research
- ✧ An infrastructure OF or FOR information?
 - ✧ OF: framework to support any kind of information
 - ✧ FOR: framework to provide context for interpretation, use, re-use



Summary-2

- ✧ e-Research for all?

- ✧ Documents

- ✧ Common need for access, curation

- ✧ Differences in variety, timeframe

- ✧ Data

- ✧ Common need for capture, access, curation, tools

- ✧ Differences in variety, timeframe, representation, sensitivity, economics, incentives, existence of repositories



Summary-3

- ✧ Context and interpretation
 - ✧ Scholarly publications provide context for research
 - ✧ Data repositories remove context from research
- ✧ Incentives to share and not to share
 - ✧ Share for common good
 - ✧ Not share for priority, competition, ownership
- ✧ Shaping a common information infrastructure
 - ✧ Base design on context, use, re-use, sharing of information
 - ✧ Identify areas with opportunities for common tools, representations, repositories
 - ✧ Access to information will depend more upon behavior, trust, and policy than on technology