Interoperability, Standards and Linked Data: promise and challenge to SDK

Jim Hendler

Tetherless World Professor of Computer and Cognitive Science
Assistant Dean of Information Technology and Web Science

Rensselaer Polytechnic Institute

http://www.cs.rpi.edu/~hendler

@jahendler (twitter)

- What can Science learn from data-sharing on the World Wide Web
- Shattering a couple of myths
  - Science is not the big kid anymore
  - Expressive ontologies are not a scalable solution
    - +1: Lightweight integration trumps ontology
    - See http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html
- Posing some new issues/approaches
  - Data search is critical
  - Usable technologies are transitioning
- Challenges remain
  - Which I'll let the next speaker address

- Yes, science uses some extremely large databases and many of them are crucial to society
- World Wide Web data is also extremely large
  - Maybe not crucial to society
- And much better "funded"
  - eg. Facebook
    - 25 Terabytes of logged data per day; valuation $33B (NIH budget ~ $31B)
  - eg. Google
    - In 2008 it was estimated at 20 petabytes per day (not including youTube); current valuation $190B (about 1/3 the ***entire*** DoD budget)

- Annotation and semantics in papers/literature search
  - Important, but really not the key *unaddressed* problem
    - Keyword search
      - Available as a service
    - Informatics/language extraction
      - Available as open source products
        - » Application costs for known technologies is a problem, but not the research problem
        - » Most researchers are trying to invent this rather than use the off the shelf solutions
    - IBM Watson shows what Q/A will look like in a few years
    - …
  - Note there's some interesting new work in embedding the annotations
    - (Ask Hal)

- Moving away from relational models
  - cf. NoSQL
  - cf. BigData
- Moving towards interoperability and exchange
  - Example: like buttons from Facebook
  - Example: Google
- **Simple metadata and lightweight semantics are becoming a big deal**
  - http://www.slideshare.net/jahendler/the-semantic-web-2010-update

Example: Government Data on the Web

Government Data Sharing

Tetherless World Constellation

January 1, 2009

"Openness will strengthen our democracy and promote efficiency and effectiveness in Government."

--- President Obama

May 21, 2009

data.gov online

57 Data Sets

December 8, 2009

"Open Government Directive" released

~2000 Data Sets

May 21, 2010

data.gov relaunch with semantic web featured

>305,000 Data Sets

2009

2010 …

June30,2009

Putting Govt Data online-
Data.gov.uk beta

January 19, 2010

data.gov.uk online

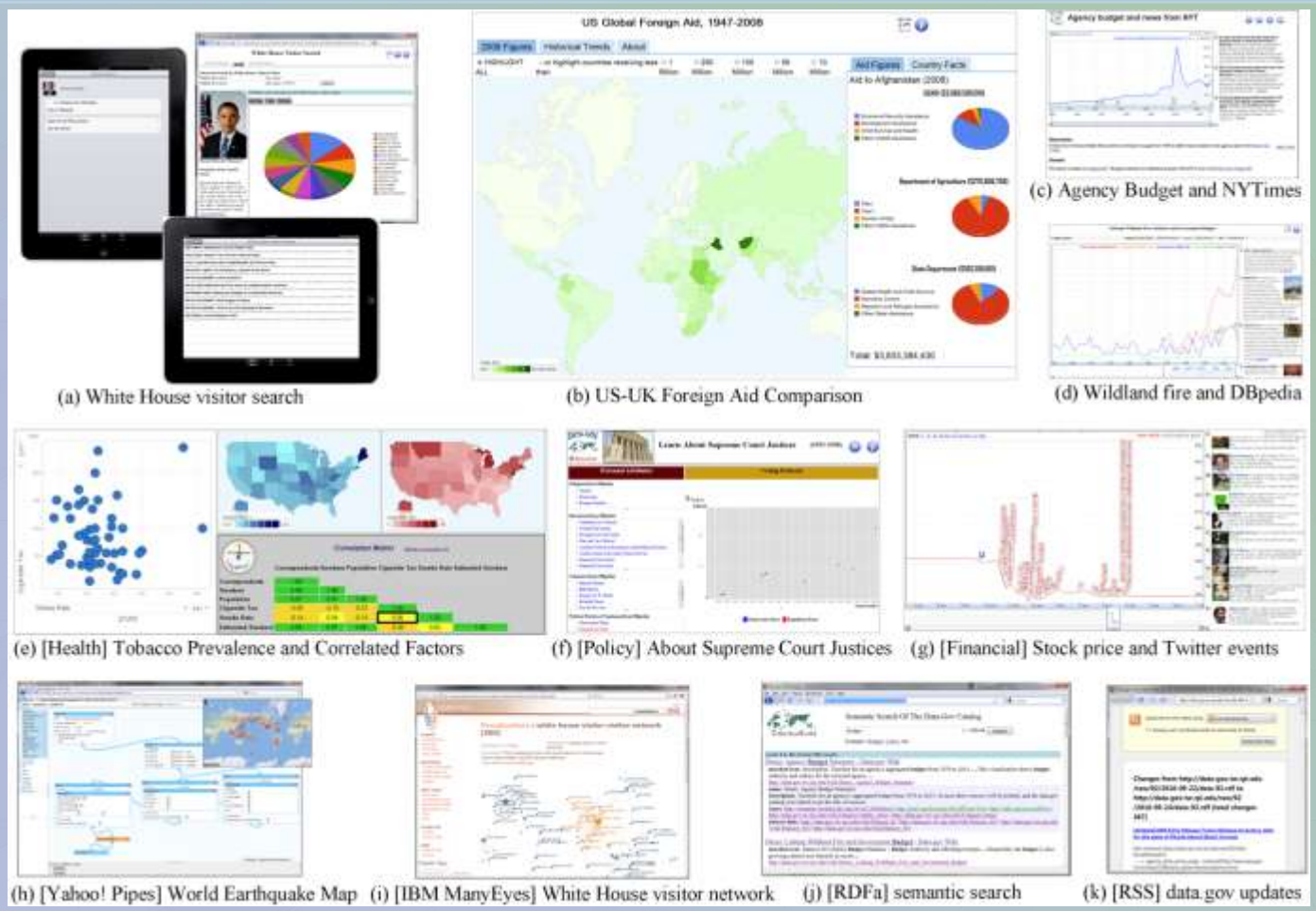
Data.gov community: International

Tetherless World Constellation

Examples:	
US	305,000
Japan	30,184
Denmark	17,086
UK	6,000
Korea	833
Australia	700
World Health Org	400
Ireland	263
Catalonia	246

Creating/Using Data "app" technologies

Tetherless World Constellation



(a) White House visitor search
(b) US-UK Foreign Aid Comparison
(c) Agency Budget and NYTimes
(d) Wildland fire and DBpedia
(e) [Health] Tobacco Prevalence and Correlated Factors
(f) [Policy] About Supreme Court Justices
(g) [Financial] Stock price and Twitter events
(h) [Yahoo! Pipes] World Earthquake Map
(i) [IBM ManyEyes] White House visitor network
(j) [RDFa] semantic search
(k) [RSS] data.gov updates

See more than 50 of these at http://logd.tw.rpi.edu

Thought: Easy to credit app downloads/usage

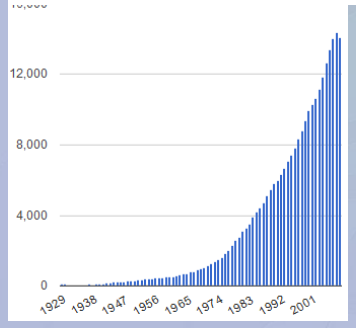
Linking GDP of the US and China

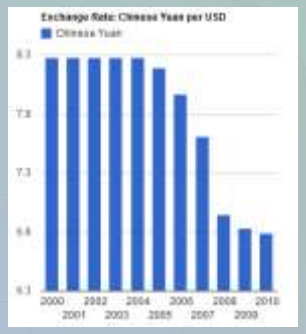
Tetherless World Constellation

BEA
BUREAU OF ECONOMIC ANALYSIS
U.S. DEPARTMENT OF COMMERCE

Federal Reserve Statistical Release
H.10
Foreign Exchange Rates (Weekly)

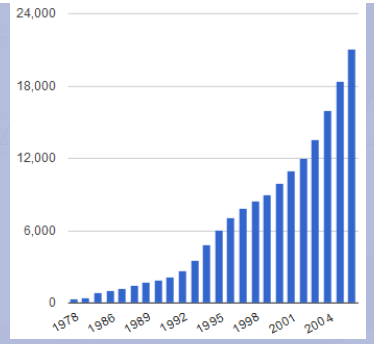
GDP of the US (Billion Dollar)



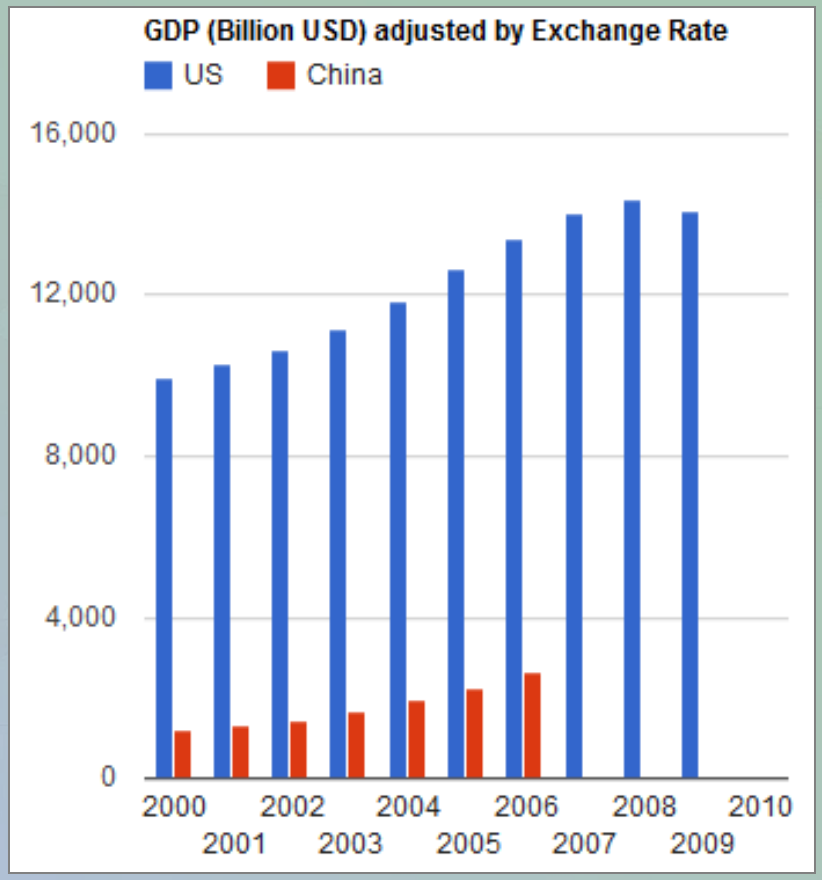


中华人民共和国国家统计局
National Bureau of Statistics of China

GDP of China (Billion Chinese Yuan)



TWC LOGD
Rensselaer



[Temporal Mashup] bea.gov + federalreserve.gov +stats.gov.cn

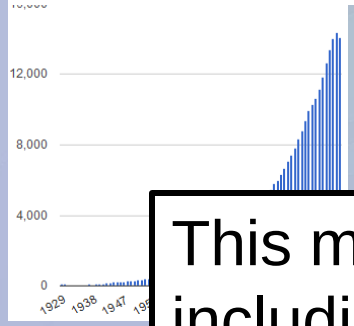
Linking GDP of the US and China

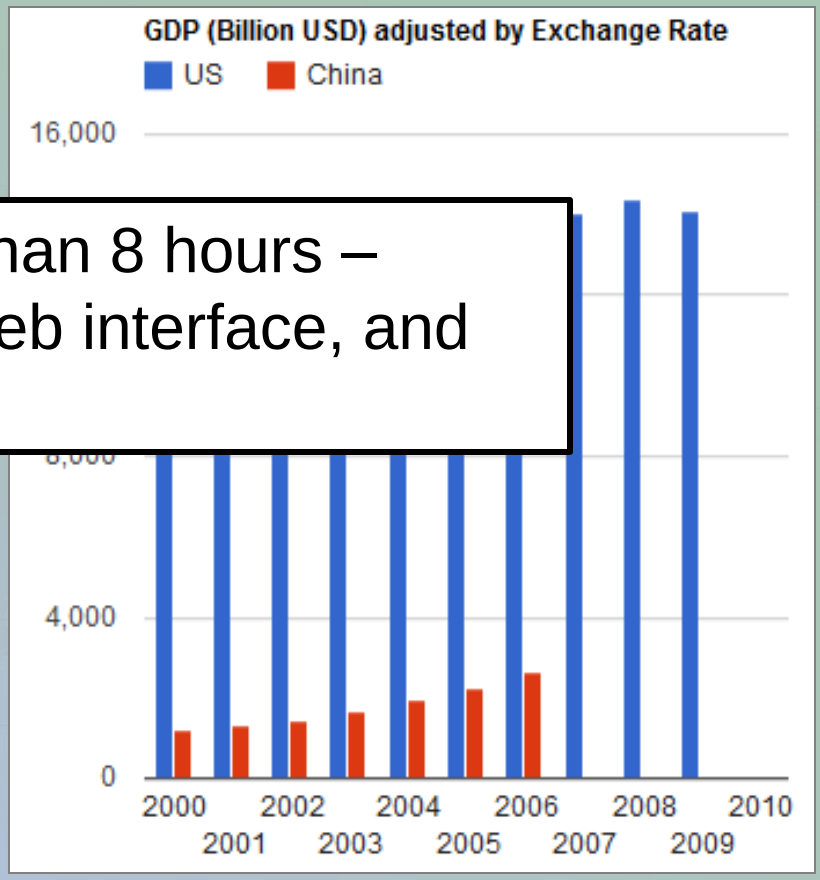
Tetherless World Constellation

BEA
BUREAU OF ECONOMIC ANALYSIS
U.S. DEPARTMENT OF COMMERCE

Federal Reserve Statistical Release
H.10
Foreign Exchange Rates (Weekly)

GDP of the US (Billion Dollar)

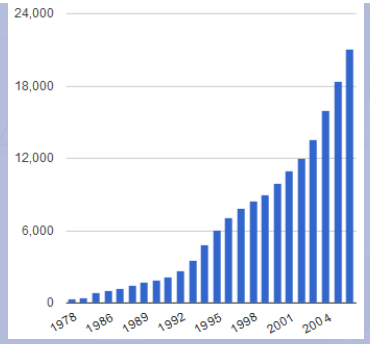




This mashup was built in less than 8 hours – including conversion of data, web interface, and visualization!

中华人民共和国国家统计局
National Bureau of Statistics of China

GDP of China (Billion Chinese Yuan)

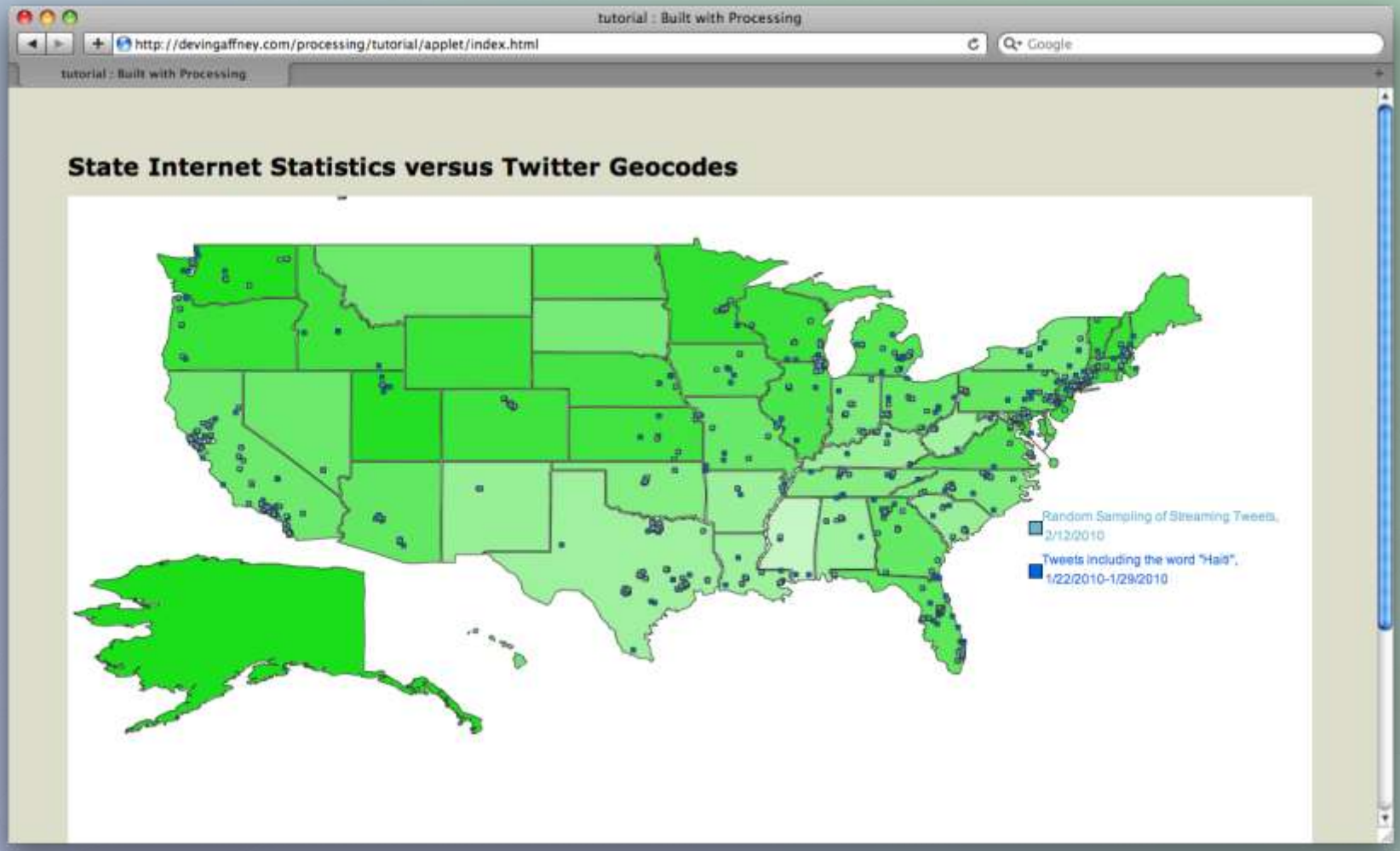


TWC LOGD
Rensselaer

[Temporal Mashup] bea.gov + federalreserve.gov +stats.gov.cn

Govt data linked to Social Media Metadata

Tetherless World Constellation



There is a lot of workflow information in the mix

Tetherless World Constellation

Convert

derive

Access

derive

Enhance

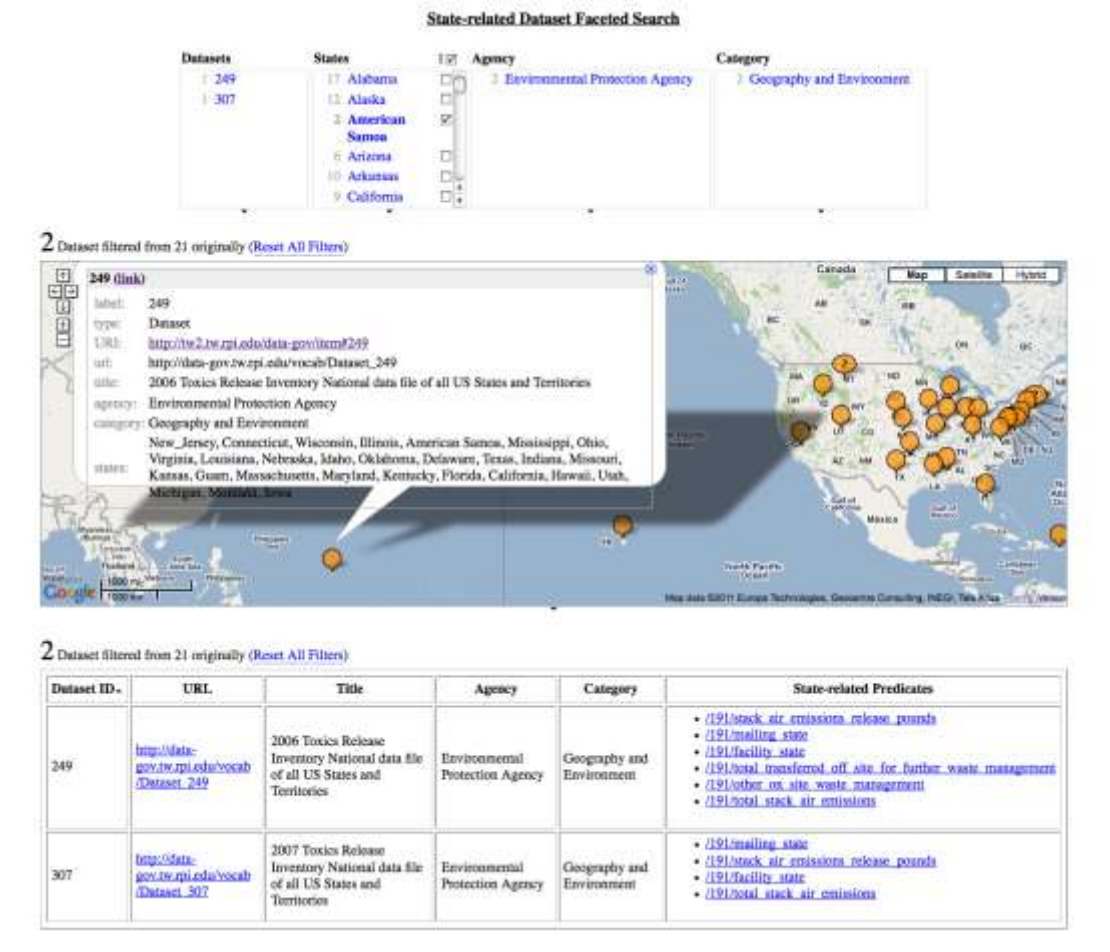
create

revision

Version

SemDiff

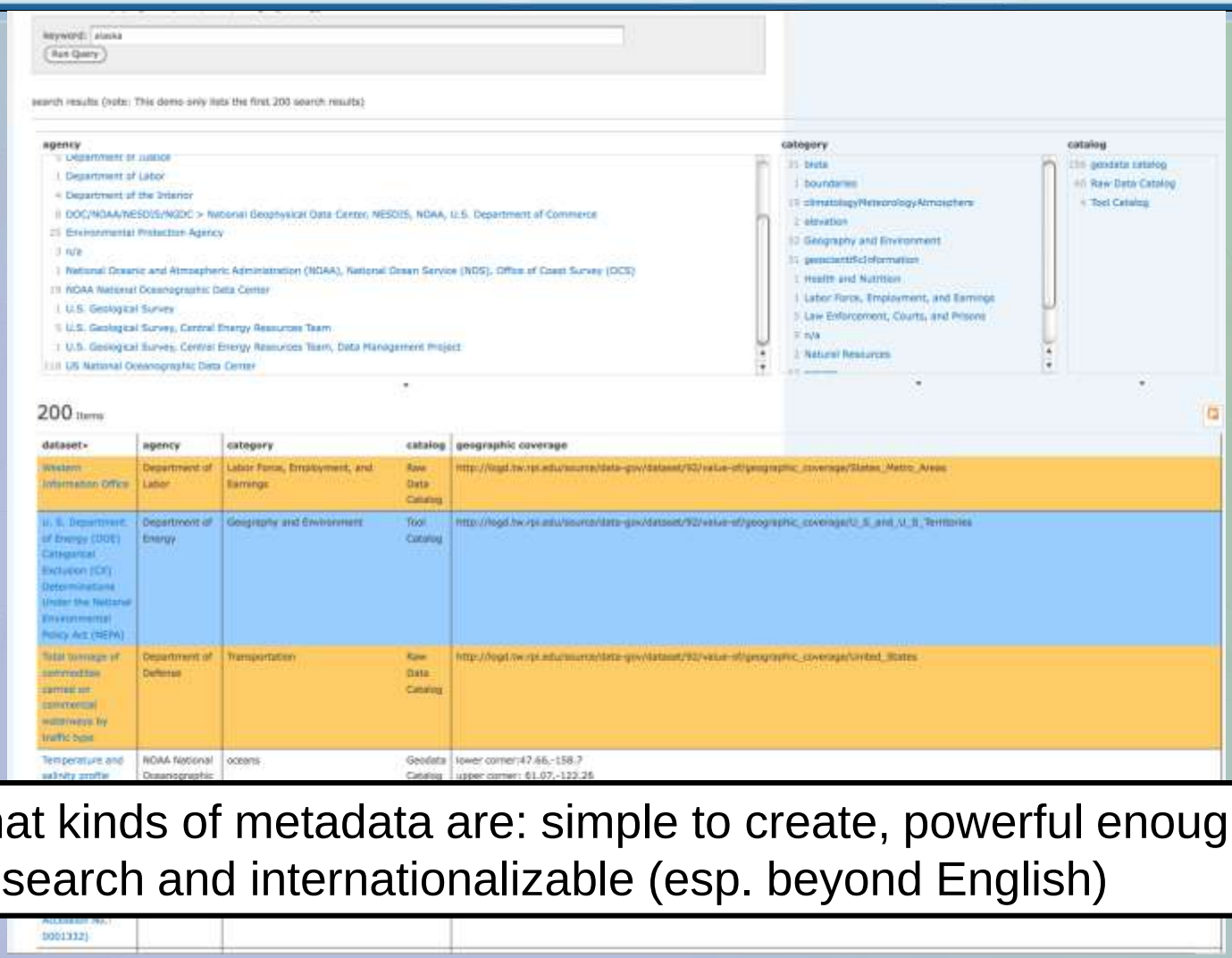
derive



How can we search for data?

Metadata is crucial

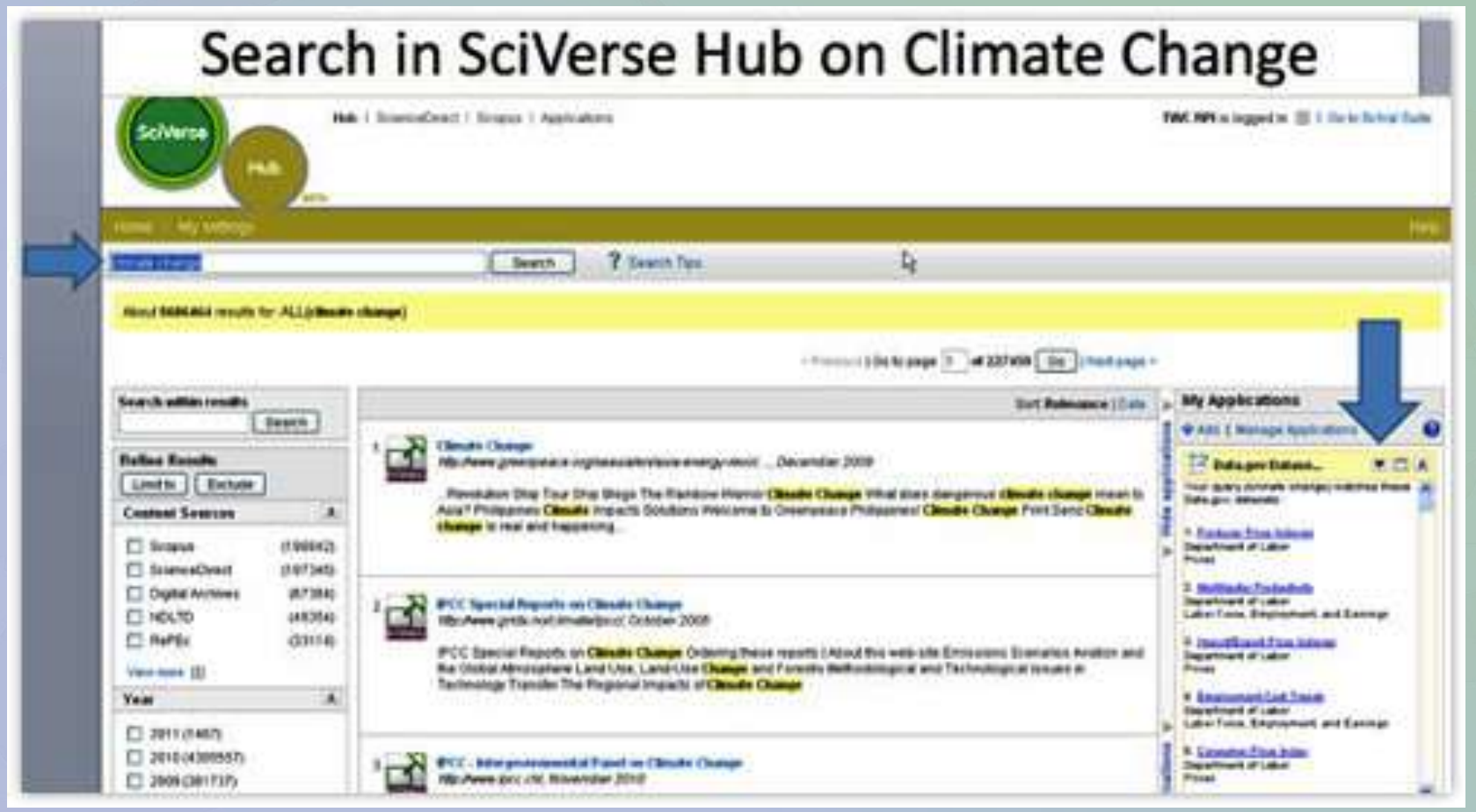
Tetherless World Constellation



What kinds of metadata are: simple to create, powerful enough for search and internationalizable (esp. beyond English)

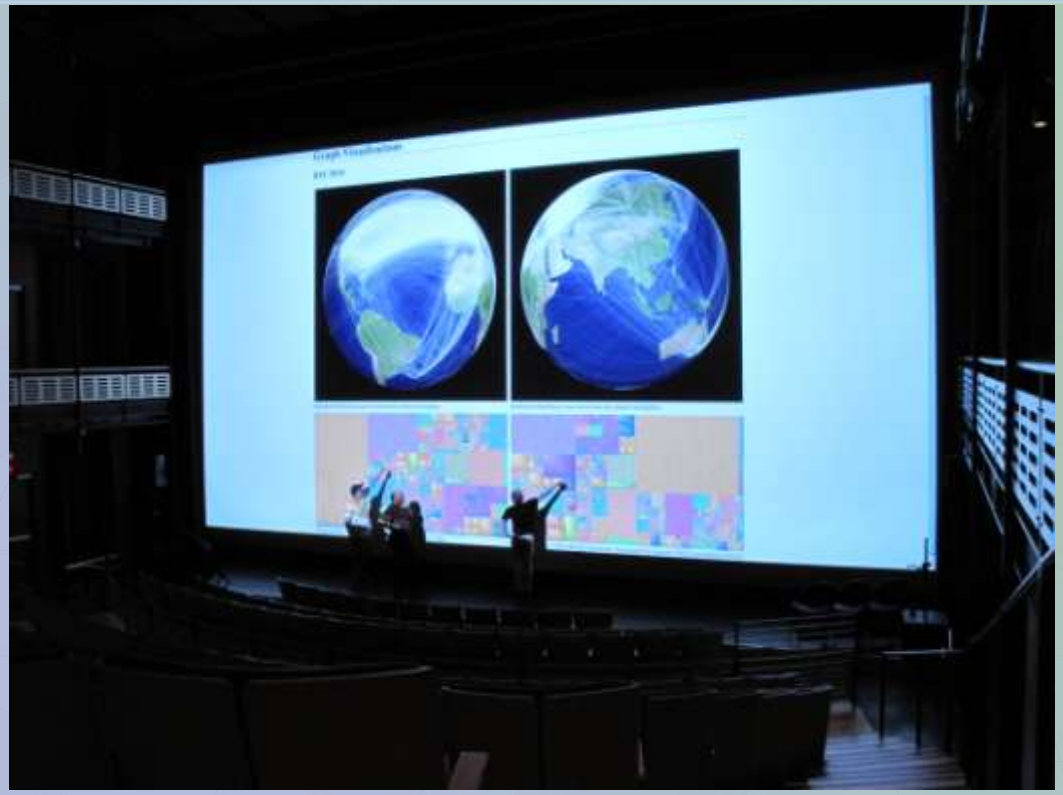
Example, integrating data and info search

Tetherless World Constellation



Challenge: Visualization can be more than art

Tetherless World Constellation



Fox &Hendler, Changing the Equation in Scientific Visualization Science, 2/11/1

How can we make data visualization a 1st class citizen in the scientific methods?

- Linked open data (cf. data.gov, data.gov.uk)
  - Being explored in Genomics, Astronomy
  - Needs a lot more attention in other scientific domains
- Markup languages and semantics and tools can enable "transparency"
  - Interdisciplinary science requires data communication
- Lower barriers to internet visualization, e.g. Google vis, MIT simile, many more…
- Web 2.0 to put people in the loop and use and contribute to annotations

- Finding Stuff!
  - How do I find a dataset in the many out there that might be of use to me?
    - Cannot keyword search in data
    - Cannot
  - How do I know what is in a large data store? In a virtual observatory? In the cloud?
    - What is the coverage?
    - What is the access?
    - Who do I need to ask for what
  - What are the rules about using it?
    - What can I combine it with?
    - How do downstream users know I've combined it

- Trust
  - Government data is controversial, and potentially biased
    - How do we confirm or dispute?
- Combination
  - When we combine data we need to keep the provenance of information (see trust)
    - How can we show and use?
- Scaling
  - Data-gov Wiki has already converted 5,448,693,510 triples
- Versioning and updating
- Archiving
- Searching
- …

ASK ME ABOUT MY

New PhD Program

RPI is considering the creation of a

SCIENCE INFORMATICS

PhD program