CouchDB, BibJSON and BKN

Schemaless databases and bibliographic metadata

Background

- Me - Nitin Borwankar
- twitter.com/nitin
- tagschema.com
- Data - RDBMS Ingres, Sybase 91-93
- Data on the web 1994 - 2004
- Data 2.0 2005 - 2008
- Data NG 2008 -...

What this is not about

- Erlang
- Concurrency
- Actors
- Pattern Matching
- Functional Languages
- anything to do with this conference ..

What this IS about

- Schemaless data
- Bibliographic metadata - and why you should care
- JSON and how it helps immensely
- CouchDB and why it is critical to BKN
- Why the CRUD metaphor is broken

First a word from our sponsors ...

- Bibliographic Knowledge Network
- NSF funded 2 yr grant started in Sept 08
- Harvard, Stanford, UC Berkeley, American Inst of Mathematics
- bibkn.org

PI Jim Pitman

- Prof UCB Dept of Stats
- Past President of IMS (Inst of Math Stats)
- Strong advocate for open data access
- Python/data hacker
- Uncompromising on problems with RDBMS

Major Efforts

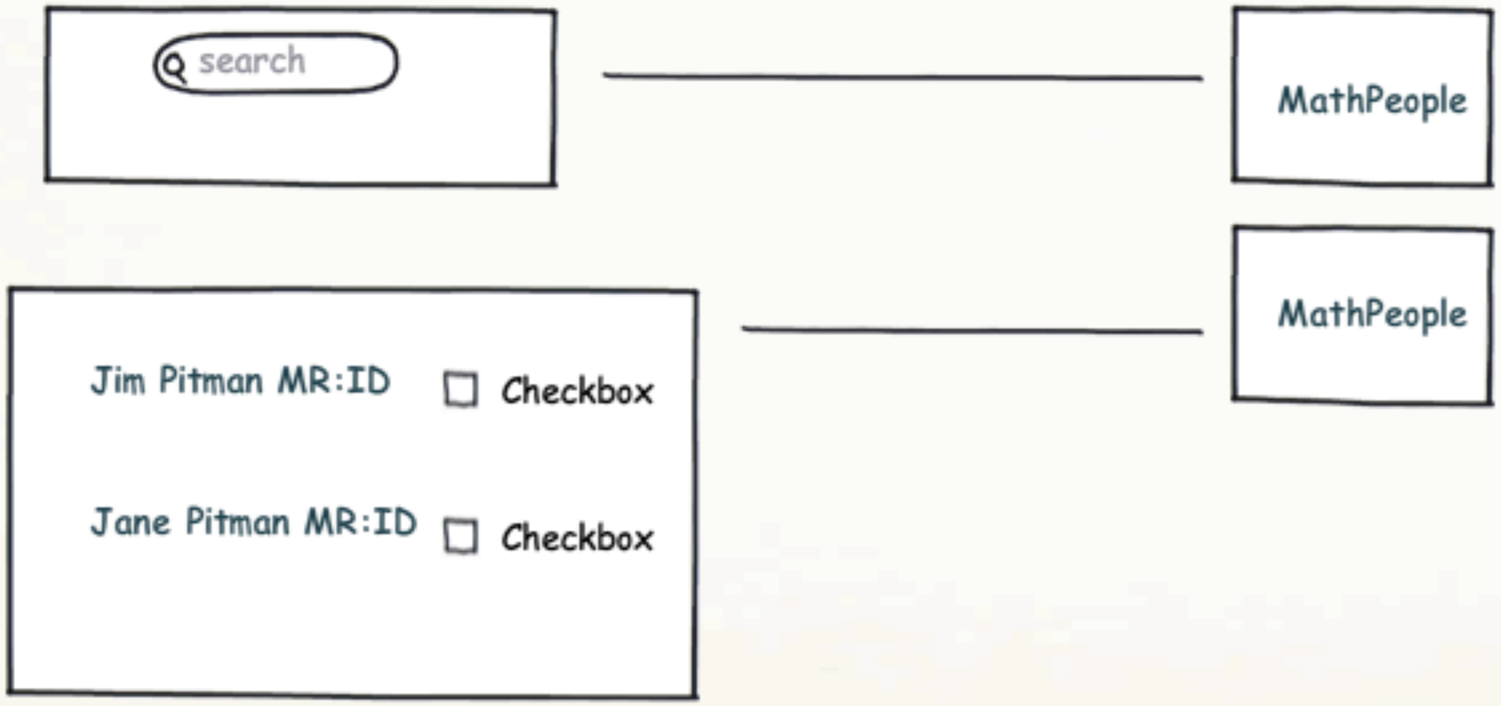
- Bibliographic analysis - citations as data
- Social networks of scientific collabrn
- Major foundation problems with implications for Web 2.0 and beyond
- Glue between Web 2.0 and Semantic Web

The problem

- Loosely typed data - bibtex publ focused
- Bibliographic primordial soup
- Name disambiguation, Subject disambiguation
- Over small well defined universe

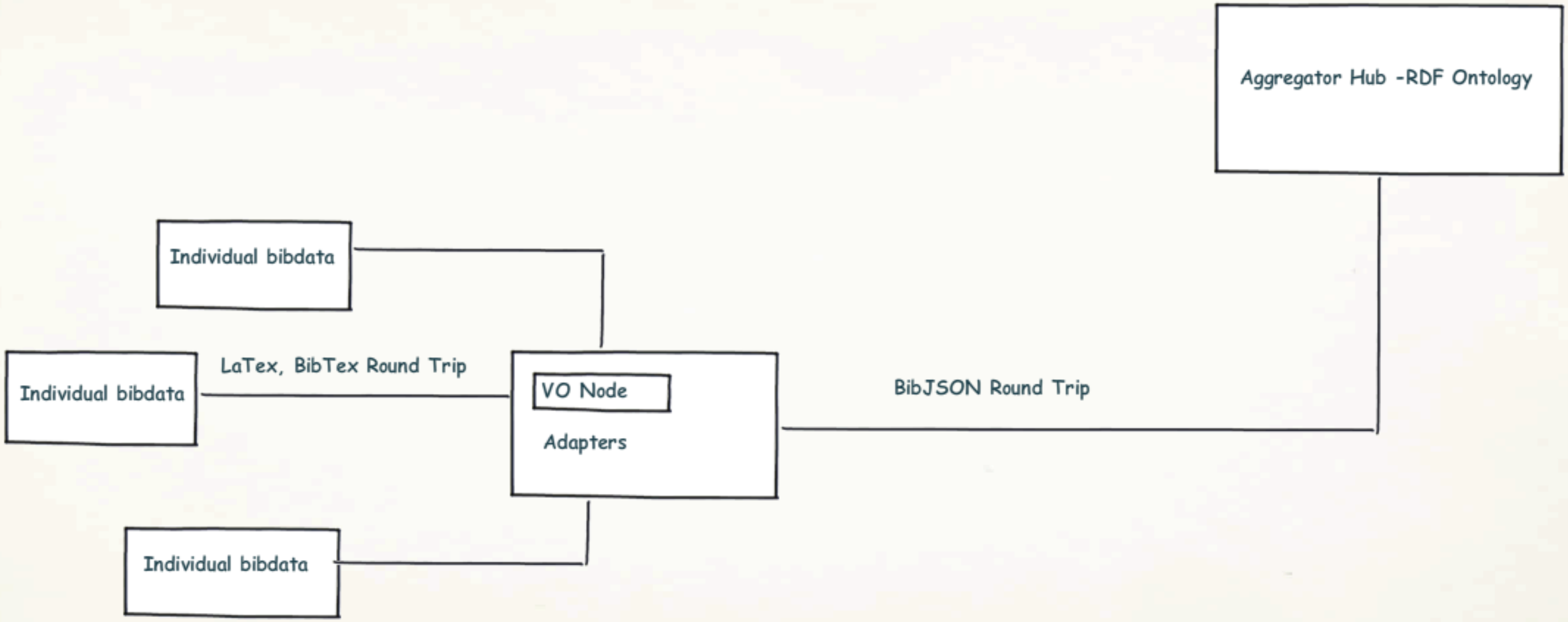
Example

- Fractional distillation of data
- Bootstrap
- Disambiguation protocol
- opaque string ----->
- <------- list of ids
- ------> disambiguated id
- Crowdsourcing data cleanup


search
MathPeople
Jim Pitman MR:ID
☐ Checkbox
Jane Pitman MR:ID
☐ Checkbox
MathPeople

Architecture

- Messy data at edge
- Cleanup via fractnl distlln
- bibJSON in CouchDB
- RDF at core


Aggregator Hub -RDF Ontology
Individual bibdata
Individual bibdata
LaTex, BibTex Round Trip
VO Node
Adapters
BibJSON Round Trip
Individual bibdata

Schemalessness is critical

- Google Contacts UI
- CRUD and CRUD-ER
- E - Extend - add attributes
- R - Restrict - remove attributes
- On an instance record basis not class

RESTfulness is valuable

- Integration with Drupal
- User management
- Permissions

Robustness is relaxing

- Problems with GAE
- Problems with AWS

Extensibility is exhilarating

- Problems with MySQL/SQLite
- fixed schema
- URL’s and ID’s

Hence bibJSON

- link to bibJSOn spec
- sample of bibJSON
- demo of Jim’s stuff

bibJSON and CouchDB

- Ubuntu 8.10 AMI on EC2
- Dependencies via apt-get
- CouchDB 0.9 built from tarball
- couchdb-python client in cgi to Apache 2.2