temporal logic

- is-jetlagged(Joe)
- doesnt-want(to-bore-you, Joe)
- concerned(Joe)
- may-need(Joe, periodic-reminder)
Data Sharing in Laboratory Science
towards a Protégé solution

Joe Edelman
formerly: fMRI Data Center (NIH/NSF)
still: Dartmouth College
Overview

- the problem of scientific data management
- our approach to it: data model & interfaces
- current status of the project
Scientists want...

- to understand & reanalyze data from outside their discipline
- to analyze someone else's data as easy as their own
- to apply new analysis techniques to large quantities of old data automatically
- to use model-agnostic techniques (clustering, etc) to mine and to relate new raw data to older data
But...

different data/knowledge exchange for lab-local and interlab use

lab local: C-structure binary files, columnar text files, spreadsheets, and form-based systems like Matlab's GUI, MS Access, and FileMaker Pro

interlab: journal articles, figures, and a variety of semi-structured networked results databases (Genbank, ACEDB, PDB, etc).

(re)analysis can only be done lab-locally or by special request
So...

Layered Solution

- make a core semantic data model for laboratory science
- deploy tools that operate on this core for exchange of data & knowledge between fields
- extend this core for each domain or even each lab, to ease lab-local use
- work towards the ease-of-use and robustness of the best existing systems: files, forms, RDBMSes
But...

Challenges

- bridging data & knowledge systems
- files are hard to beat
  - easy to understand, organize, and integrate
- FileMaker is hard to beat
  - high data integrity, simple forms
- extensibility can be messy
  - reconciling parallel extensions
  - maintaining a useful core
Try anyway

- tell you about the data model
- tell you about the interfaces
- review these challenges to see how well we've done
An eScience Ontology
Using just these core relations, we can display a graphical interdisciplinary summary of the experiment.

Scientists have been very positive about this summary in interviews.
Graphical Depiction
(a key advantage)

input-to, output-from, DIRECT-TYPE, consists-of, performed-by, model

Using just these core relations, we can display a graphical interdisciplinary summary of the experiment.

scientists have been very positive about this summary in interviews
Experimental Design, Setup, Recording, and Analysis Ontology (EDSRA)

1. **Researcher**
   - **Subject**
   - **Experimental**
     - **Person**
     - **Machine**
     - **Measurement**
     - **Theory**
     - **Product**

2. **Waveform**
   - **Image**
   - **Date**
   - **Array**
   - **Volume**
   - **Successive Images**
   - **Successive Volumes**

3. **Grant**
   - **Experiment**
   - **Trial**
   - **Happening**

4. **Permutation**
   - **Production Process**
   - **Design Process**
   - **Recording Process**

5. **Experiment**
   - **Design Process**
   - **Recording Process**

6. **Macroevent**

---

**Experimental Design**

**Setup**

**Recording**

**Analysis**

**Ontology** (EDSRA)
EDSRA-fMRI

Diagram:
- **Experiment**
  - fMRI Experiment
  - Machine
    - fMRI Scanner
    - Video Display
  - Theory
    - Attentional Theory
  - Timecourse Condition
    - Waveform
    - Volume
      - Anatomical Scan
    - Anatomical Scanfile
    - fMRI Scanfile

- **Trial**
  - Run

- **Recording Process**
  - Anatomical Scan
  - fMRI Scan
EDSRA - Queries

measurement / recording
- what is being recorded?
- what device was used to make the measurement?
- is this measurement continuous or discrete?
- etc

analysis
- has this data been normalized, filtered, etc?
- what analyses are available? which possible analyses apply?
- can this analysis be reproduced exactly from its antecedents?
- etc
EDSRA - Decompositions

- as various experiments
- as various kinds of objects about which data has been collected (human subjects, tasks, scanners)
- as measurements, volumes, waveforms, etc.
- as tested mathematical models and theories
Method of Extension

- facet overrides used extensively
- keeps all information accessible using core model
- provides for domain specific guidance in data entry
interfaces to the ontology
Advantages of Protégé

- preeminent tool for managing data in semantic data models
- cleanly designed in Java with easy extension in mind
- multiplatform, can be RDBMS or file backed
- guides data entry / knowledge acquisition intelligently using ontology
- can be queried in many ways by using different tabs
- thriving community
lacking in Protégé

capability
- support for large multihomed files
- support for measurements, arrays, mathematical functions, etc

usability
- simple hierarchical and graphical methods of browsing (comparable to those for filesystems)
- alternatives to KnowledgeBase API java for getting data in and out programmatically
- support for derived data / views / normalization
our extensions

- support for large multihomed files
- support for measurements, arrays, mathematical functions, etc
- simple hierarchical and graphical methods of browsing (comparable to those for filesystems)
- alternatives to KnowledgeBase API java for getting data in and out programmatically
- support for derived data (and normalization)

Coming this fall
support for large multihomed files

- UUIDs / surrogates
- checked against multiple file resolvers
- resolved to URLs
- good with NFS, AFS, HTTPS, etc
- "tfile://" prefix supports removable media

work by Jeff Woodward
support for measurements, arrays, mathematical functions

- units of measure: attribute of “float” scalar and compound (recording, image, volume) types
- quantity: ontology and data entry support, no query support
- 1D, 3D, and 4D visualization
support for measurements, arrays, mathematical functions

- **units of measure**: attribute of “float” scalar and compound (recording, image, volume) types

- **quantity**: ontology and data entry support, no query support

- **1D, 3D, and 4D visualization**
support for measurements, arrays, mathematical functions

- **units of measure**: attribute of “float” scalar and compound (recording, image, volume) types

- **quantity**: ontology and data entry support, no query support

- **1D, 3D, and 4D visualization**
simple hierarchical and graphical methods of browsing
(comparable to those for filesystems)

- ontology-neutral
- knowledge explorer
- browser formats
- coalescing graph widget

Coming this fall
Explorer Tab

- Toolbar knows context
- Automatic or manual root cls
- !Contextual menus
- !Avoids proliferation of instance windows
Browser Formats

s = "Assessed ${:DIRECT-TYPE}"  
+ " of ${assessment-subject}"  
+ " on ${start-date}";

project.setDirectBrowserFormat(assessmentCls, s);
Coalescing Graph Widget

- A. Petr Jenata
- B. 21 Human Subjects
- C. Siemens Scanner
- D. Video Display
- E. E Prime software
alternatives to KB API for programmatic data access

- Two simpler syntaxes:
  - Simple Java API
  - Component Paths
- Perl, Python, & Unix command line bindings
- FormWidget Actions

```
listInstances("Dog")
  .with("Owner", i);

dog.list("owners")
  .makeWith("name", "Jim");
```
alternatives to KB API for programmatic data access

- Two simpler syntaxes:
  - Simple Java Queries
  - Component Paths

- Perl, Python, & Unix command line bindings

- FormWidget Actions

"/NSF2003/Motion/Scan:jxe-sep28-1/output"
alternatives to KB API for programmatic data access

- Two simpler syntaxes:
  - Simple Java Queries
  - Component Paths

- Perl, Python, & Unix command line bindings

unix% wsrun http://fmridc.org/jws/daily/JavaServer.jnlp

unix% protege-add "/NSF2003/Motion/Scan:jxe-sep28-1/output" scan.img
this fall

solves problems with reified relations by making them appear as simple slots at the API level

will allow viewing our KB as “just products” or “just processes”
Again...

Challenges

- bridging data & knowledge systems
- files are hard to beat
  - easy to understand, organize, and integrate
- FileMaker is hard to beat
  - high data integrity, simple forms
- extensibility can be messy
  - reconciling parallel extensions
  - maintaining a useful core
current status of the project
50 neuroimaging experiments encoded using the ontology, available at http://fmridc.org

3 meta-analyses in the works

deploying as lab-local data management solution at Berkeley, MIT, and Washington University Radiology Labs.
future development of some features in some doubt

all extensions open source and available

preprint of ontology paper available soon

http://fmridc.org/dmt
http://sf.net/projects/fmri-dmt

jxe@dartmouth.edu
## Availability

http://sourceforge.net/projects/fmri-dmt

| Laboratory Features |  
|---------------------|---
| files               | /files  
| measurements & dates | /widgets, /units  
| 1D viewer           | /timecourse  
| 3D/4D viewer        | /viewer, /image  

| User Interface |  
|----------------|---
| explorer tab   | /ke  
| browser formats| jxe@dartmouth.edu  
| coalescing graph | 12/2003  

| Programmatic Access |  
|---------------------|---
| simple java api     | /souffle/src/org/fmridc/protege/ProtegeUtilities.java  
| pathname api        | jxe@dartmouth.edu  
| perl bindings       | /souffle/scripts/putscan.pl  
| unix bindings       | /souffle/scripts/putscan.pl  
| virtual slot framestore | 12/2003  
| python bindings     | jnw@dartmouth.edu  

Laboratory Features:
- files
- measurements & dates
- 1D viewer
- 3D/4D viewer

User Interface:
- explorer tab
- browser formats
- coalescing graph

Programmatic Access:
- simple java api
- pathname api
- perl bindings
- unix bindings
- virtual slot framestore
- python bindings
Thanks!

There is a long tradition of outstanding open source software developed by communities of academics:

- Unix
- Emacs
- TCP/IP

It takes skilled developers to make an extensible system.

It takes an army to make a general purpose tool.