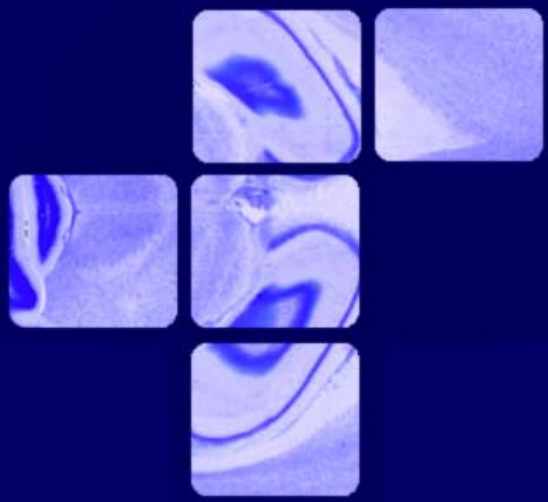

BIRN

BIOMEDICAL INFORMATICS RESEARCH NETWORK

Data Provenance

Nicole Aucoin

National Center for Research Resources

National Institutes of Health

March 3, 2005

mBIRN All Hands Meeting

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Outline

- Introduction
- Current state of the project
  - Demo
- Integration with SRB
- Integration with HID
- Instrumentation of Upload Scripts
- Requirements Gathering
- Work plans and deliverables

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Introduction

- Data provenance is…
  - Tracking what changes the data
- Data provenance is good for…
  - Recreating research for validation
  - Testing processes on new data sets
  - Information recovery when changing formats

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Current State of the Project

- Processing tools have been updated
  - Slicer, FreeSurfer, shape analysis tools
- Document type definition created to specify xml output for upload
- Raw text output can be parsed into xml files
- Documentation moved onto the wiki
  - www.na-mic.org/Wiki/index.php/Data_Provenance:Documentation

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Data Provenance DTD

```
< !-- dataprov.dtd -->
<!ELEMENT process (processStep)+>
<!ELEMENT processStep (programName, programArguments, version,
    timeStamp, (cvs)?, user, machine, platform, (compiler)?, (library)*)>
<!ELEMENT program (#PCDATA)>
        <!ATTLIST program version PCDATA #REQUIRED>
        <!ATTLIST program arguments PCDATA #REQUIRED>
<!ELEMENT timeStamp (#PCDATA)>
<!ELEMENT cvs (#PCDATA)>
<!ELEMENT user (#PCDATA)>
<!ELEMENT machine (#PCDATA)>
<!ELEMENT platform (#PCDATA)>
        <!ATTLIST platform version PCDATA #REQUIRED>
<!ELEMENT compiler (#PCDATA)>
        <!ATTLIST compiler version PCDATA #REQUIRED>
<!ELEMENT library (#PCDATA)>
        <!ATTLIST library version PCDATA #REQUIRED>
```

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Demo

- Converting a volume file from Freesurfer MGH format to COR format, viewing it in Slicer
- Raw data provenance information is captured in a text file, amidst processing output
- The raw data file is parsed and an xml file is produced

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Sample Output File

```
<?xml version="1.0" standalone='no'?>
<!DOCTYPE process SYSTEM "dataprov.dtd">

<process>
 <processStep>
  <program version="" arguments="-all-info {}">mri_convert</program>
  <compiler version="29600">GCC</compiler>
  <timeStamp>05/02/08-14:47:44-GMT</timeStamp>
  <cvs>{$Id:} mri_convert.c,v 1.84 2004/02/09 21:45:55 fischl Exp {$}</cvs>
  <machine>slicerl.bwh.harvard.edu</machine>
  <platform version="2.4.18-17.7.xsmp">Linux</platform>
  <user>nicole</user>
 </processStep>
```

Sample Output File con't

```
<processStep>
 <program version="2.5-dev" arguments="--no-tkcon --all-info --load-freesurfer-
    volume /home/nicole/data/flash40cor/COR-.info">slicer2-linux-x86</program>
 <compiler version="29600">GCC</compiler>
 <library version="4.4.2">VTK</library>
 <library version="8.4.6">TCL</library>
 <library version="8.4.6">TK</library>
 <library version="1.9">ITK</library>
 <timeStamp>02/08/05-09:48:08-EST</timeStamp>
 <cvs>\{Id: Go.tcl,v 1.81.2.2 2004/12/02 23:46:19 nicole Exp\} {}</cvs>
 <machine>i686</machine>
 <platform version="2.4.18-17.7.xsmp {}">Linux</platform>
 <user>nicole</user>
</processStep>
</process>
```

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Demo details

- Call convertandshow script, piping output to a raw file
  - Call mri_convert with –all-info flag
  - Use mri_convert to convert MGH to COR
  - Load COR volume into Slicer, with –all-info flag
- Call dataprov tcl script on raw file, piping output to an xml file

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Integration with SRB

- Upload of xml files via S commands
- Associate xml files with derived data they are describing

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Integration with HID

- Use XML parsing tools to extract the information from the XML file
  - XSLT/DOM/XPATH
- Upload values to the appropriate places in the HID
  - Integrate with the new schema

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Instrumentation of Upload Processes

- Scripts versus pipelines
- BIRNDUP
- fBIRN uploads

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Requirements Gathering

- From the Database group
  - Location to upload xml file
  - How to integrate with the HID
  - How to query from HID on data provenance fields
- From Developers
  - Keep me informed when new tools are used
    - Survey will be sent out to BIRN sites every few months
  - Help testing out information gathering and parsing on various operating systems
  - Integration of query by HID into Portal

Work Plans and Deliverables

October 2005

- Data Provenance information in SRB and HID
  - Test upload to SRB (March)
  - Give DB working group formal request for new fields (March)
  - Integrate into upload pipelines (April)
  - Test out various xml parsers (June)
  - Test upload to HID (September)
- Query by Provenance in HID (October)
- Add fields to DTD (compiler flags, data URI, data id, ?), and update the specification table (March)
- Convert DTD to a schema (March)
- (share tools/information with fBIRN) (ongoing)

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Work Plans and Deliverables

March 2006

- Query by Provenance in HID via Portal (December)
- Wrappers for third party programs (January)
  - Matlab mex files
  - SPM
  - FSL
  - Contact vendors to obtain more information, ask them to add it
- First pass on a data provenance toolkit (March)