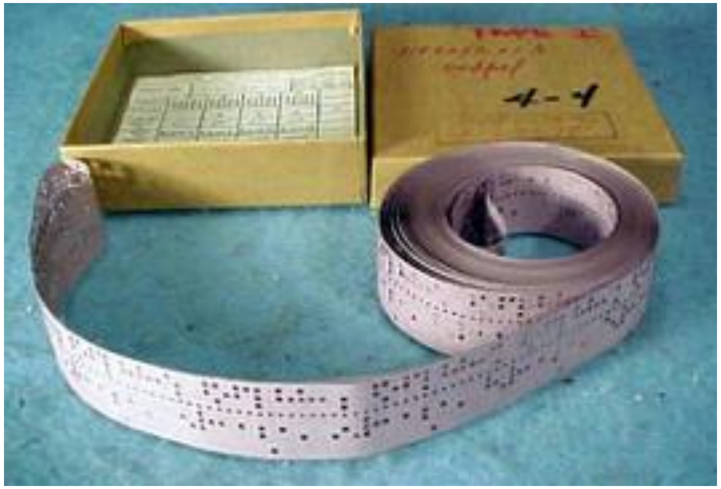
A General Purpose Toolset For Representing Data Relationships:
Converting Data Into Knowledge



Joshua Stillerman, Thomas Fredian, Martin Greenwald, John Wright

MIT Plasma Science and Fusion Center

Data Challenges

Situation

- Collecting data has never been easier
- Making sense of data – extracting knowledge – is getting harder
- Scientists are struggling to keep up with the growth In data volume and complexity

Our Thesis

- The challenge is all about putting the data into context
- **"navigational metadata"** – Context is about metadata and relationships among data objects
- In general, our approach to capturing and exploiting this class of metadata has been ad hoc and inadequate

What Sorts Of Data Might Exist From A Typical Experiment?

- Hierarchical data stores with raw and processed data
- Relational databases with “high level” results
- Electronic logbooks & annotation
- Data provenance systems
- Data catalogs
- Data dictionaries
- Information about experimental campaigns & plans
- Information about people
- Experimental proposals
- Simulation inputs & outputs
- Source code management systems
- Facility information, with details of experiment, measurement systems
- Document management systems
- Publications & presentations

Understanding Data is About Context

- In the past when things were smaller and simpler, we could keep data context in our heads
  - or in our colleague's heads
- Context is metadata about its relationships between data
- These relationships enable data discovery.
  - Adjacency to find descriptive metadata
  - Adjacency to find other interesting data
- These problems exist in almost all data intensive areas of research.
- We each build a set of ad-hoc, domain specific tools to store, explore, and retrieve this relationship metadata.
- Our team is starting to build general purpose software to address these needs.

Progressive Process of Generalization and Abstraction

Hand recorded data:

Wow – "I don't have to draw it by hand!"

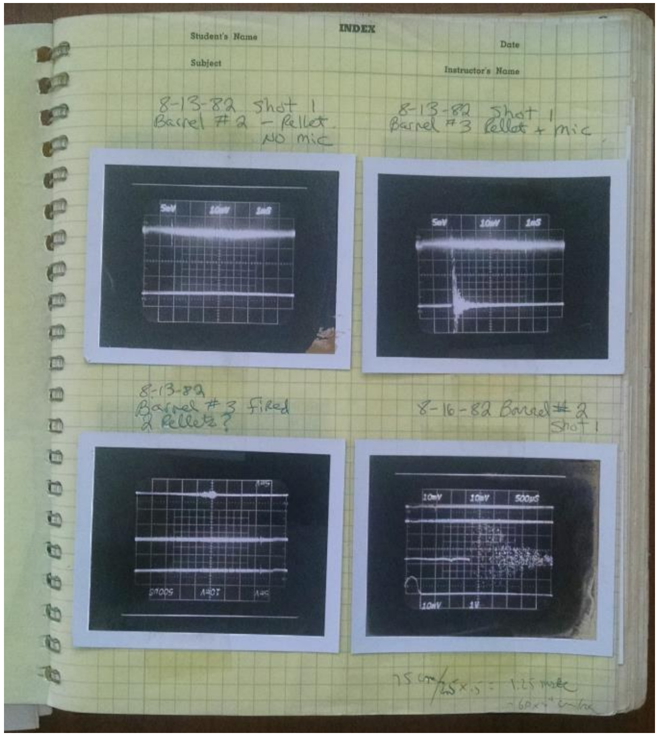
Progressive process of generalization and abstraction

Polaroid photos of oscilloscopes:

Polaroid DS34 Direct Screen Instant Camera

Progressive process of generalization and abstraction

Pasted into lab notebooks



Wow – “I don’t need to draw a picture of the screen!”

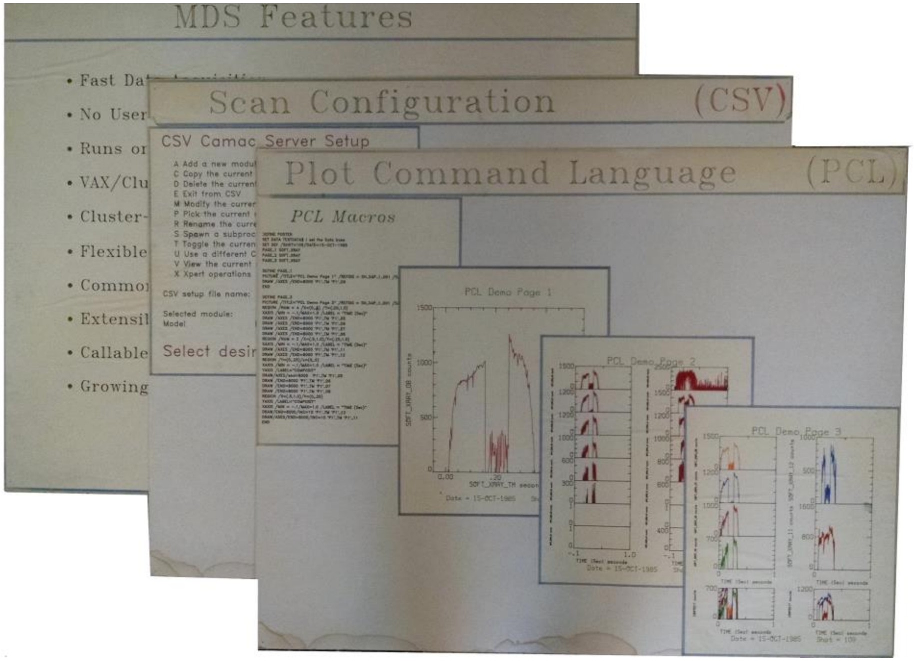
Progressive Process of Generalization and Abstraction

Purpose built data acquisition programs:

Wow – “I don’t need a ruler, I don’t need to type in the numbers”

Progressive Process of Generalization and Abstraction

Data Acquisition systems – like MDS:



MDS Features

- Fast Da
- No User
- Runs or
- VAX/Clu
- Cluster-
- Flexible
- Commo
- Extensil
- Callable
- Growing

Scan Configuration (CSV)

CSV Camac Server Setup

A Add a new modul
C Copy the current
D Delete the current
E Exit from CSV
M Modify the currer
P Pick the current
R Rename the curre
S Spawn a subproc
T Toggle the curren
U Use a different C
V View the current
X Xpert operations

CSV setup file name:

Selected module:
Model

Select desir

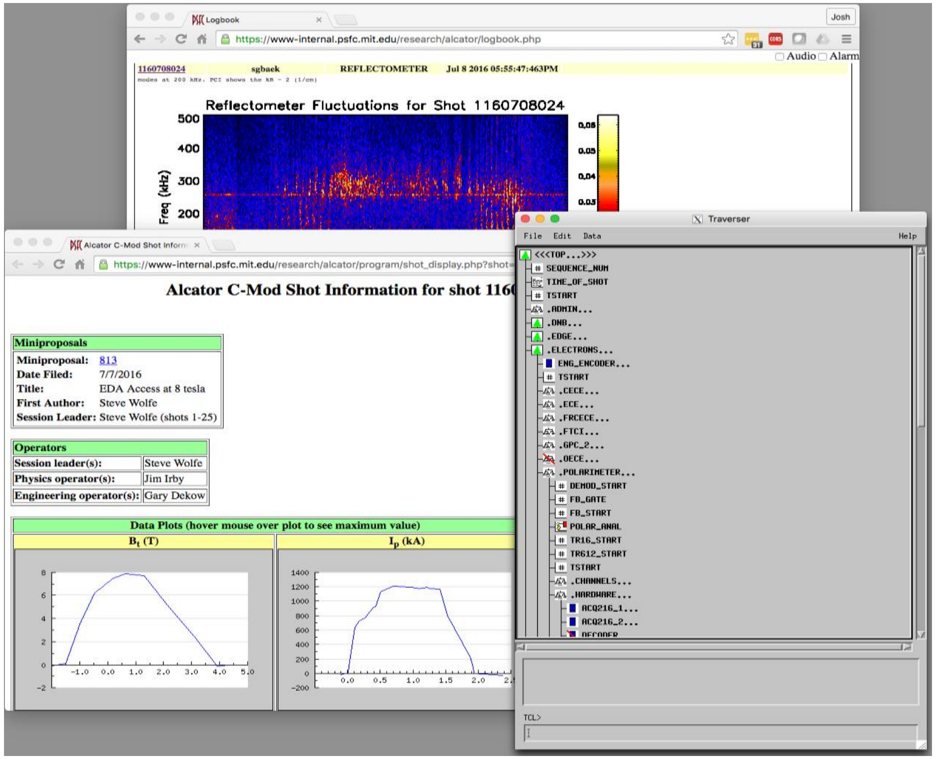
Plot Command Language (PCL)

PCL Macros

Wow – “I don’t need a programmer to get my data!”

Progressive Process of Generalization and Abstraction

Data management systems – MDSplus:



Logbook

https://www-internal.psfc.mit.edu/research/alcator/logbook.php

1160708024	sgbaek	REFLECTOMETER	Jul 8 2016 05:55:47:463PM

modes at 200 kHz. PCI shows the kR ~ 2 (1/cm)

Reflectometer Fluctuations for Shot 1160708024

Alcator C-Mod Shot Inform

https://www-internal.psfc.mit.edu/research/alcator/program/shot_display.php?shot=

Alcator C-Mod Shot Information for shot 116

Miniproposals	
Miniproposal:	813
Date Filed:	7/7/2016
Title:	EDA Access at 8 tesla
First Author:	Steve Wolfe
Session Leader:	Steve Wolfe (shots 1-25)

Operators	
Session leader(s):	Steve Wolfe
Physics operator(s):	Jim Irby
Engineering operator(s):	Gary Dekow

Data Plots (hover mouse over plot to see maximum value)

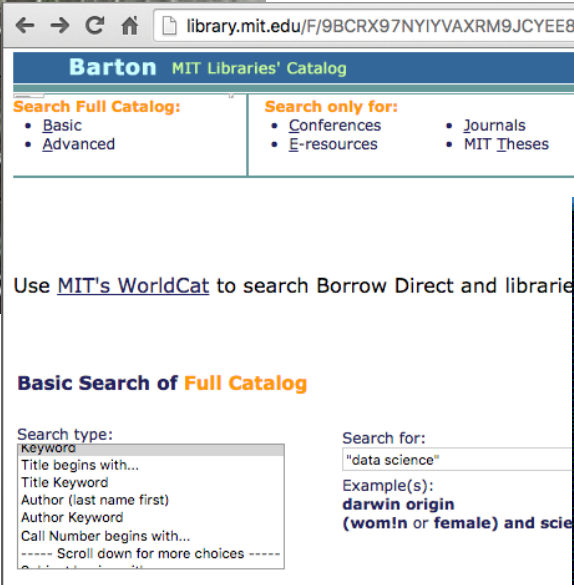
Traverser

```
<<<TOP...>>>
 SEQUENCE_NUM
 TIME_OF_SHOT
 TSTART
 .ADMIN...
 .DNB...
 .EDGE...
 .ELECTRONS...
  ENG_ENCODER...
  TSTART
  .CECE...
  .ECE...
  .FRCECE...
  .FTCI...
  .GPC_2...
  .OECE...
  .POLARIMETER...
   DEMOD_START
   FB_GATE
   FB_START
   POLAR_ANAL
   TR16_START
   TR612_START
   TSTART
   .CHANNELS...
   .HARDWARE...
    ACQ216_1...
    ACQ216_2...
    DECODER
```

Wow – “I can find out the context of this measurement!”

Progressive Process of Generalization and Abstraction

- Each step of this progression made the collection, and then organization, of collected data easier.
- When it was hard to collect data, collecting it was good.
- As it was easy to collect data, the need for organizing metadata became apparent.
- But the data still had ONE primary organization
  - Statically defined by the system implementers



library.mit.edu/F/9BCRX97NYIYVAXRM9JCYEE84CB9ERDCENL2VEBSE53AE3R2U82-00273?RN=3331...

Barton MIT Libraries' Catalog

MIT Libraries

Search Full Catalog:
- Basic
- Advanced

Search only for:
- Conferences
- E-resources
- Journals
- MIT Theses
- Reserves
- more...

- Your Account
- Help with Your Account
- Your Bookshelf
- Previous Searches

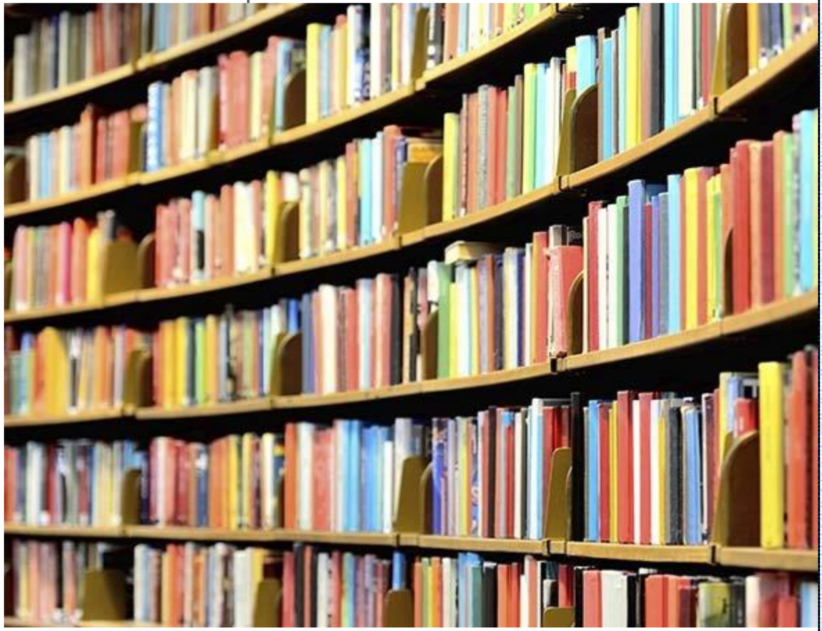
Use MIT's WorldCat to search Borrow Direct and librarie

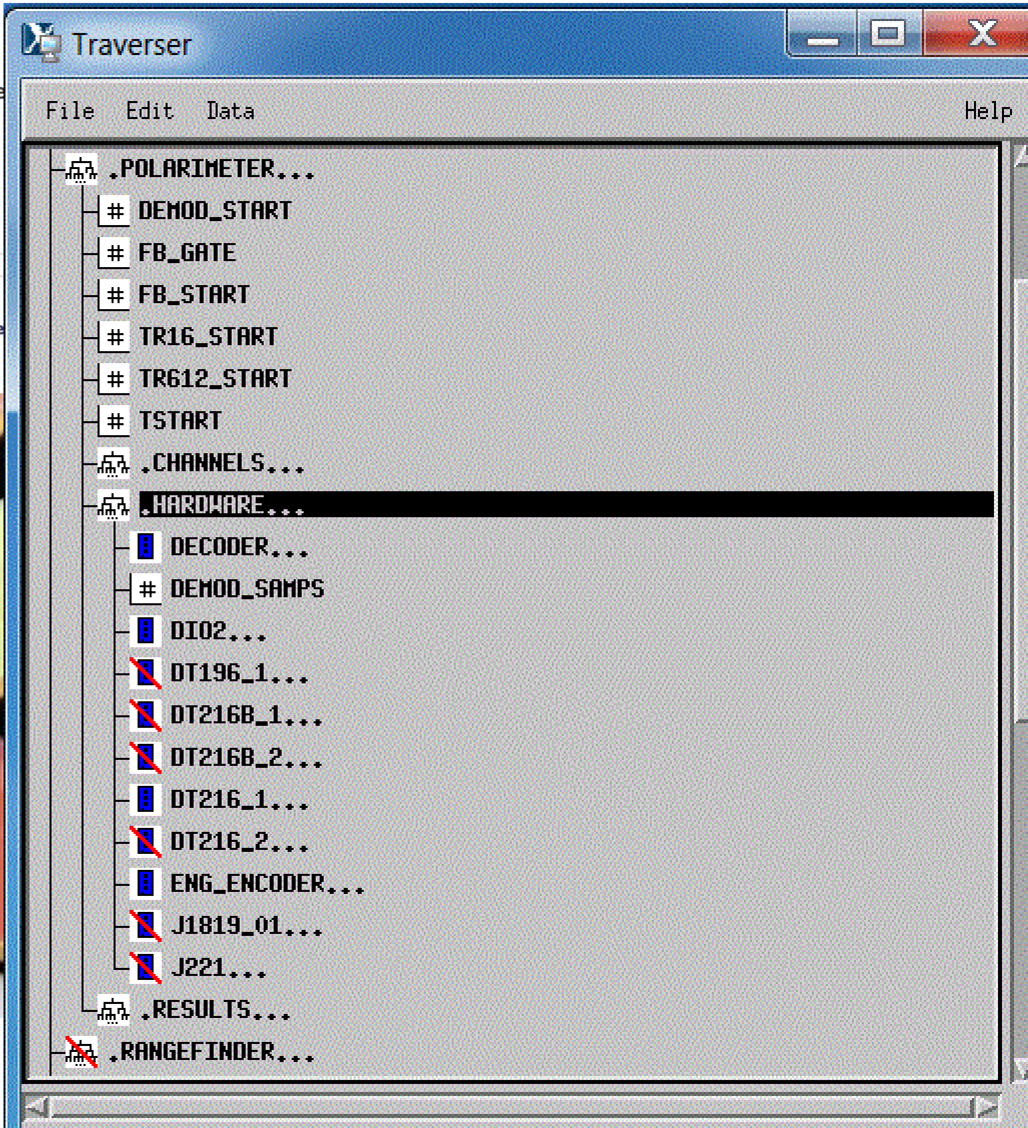
Basic Search of Full Catalog

Search type:
Keyword
Title begins with...
Title Keyword
Author (last name first)
Author Keyword
Call Number begins with...
----- Scroll down for more choices -----

Search for:
"data science"

Example(s):
darwin origin
(wom!n or **female) and scie**

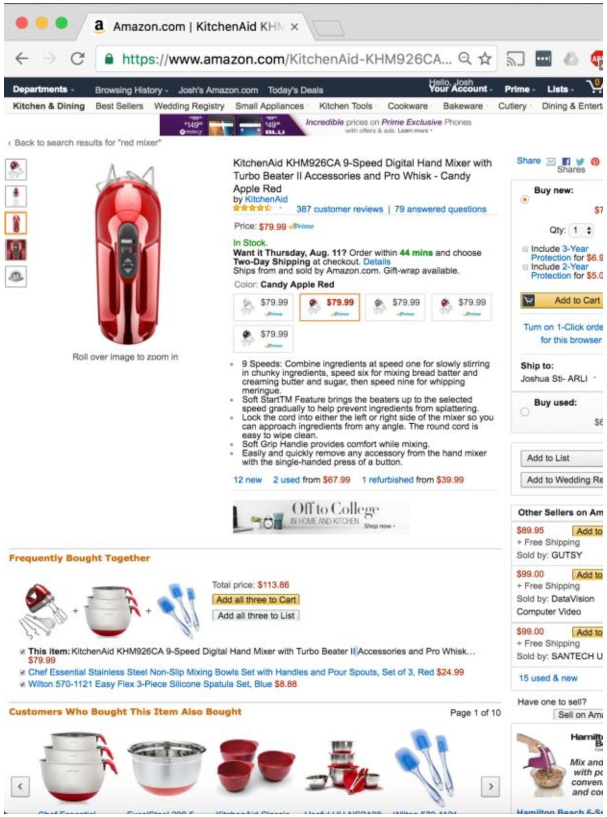




Traverser

```
.POLARIMETER...
   DEMOD_START
   FB_GATE
   FB_START
   TR16_START
   TR612_START
   TSTART
   .CHANNELS...
   .HARDWARE...
      DECODER...
      DEMOD_SAMPS
      DIO2...
      DT196_1...
      DT216B_1...
      DT216B_2...
      DT216_1...
      DT216_2...
      ENG_ENCODER...
      J1819_01...
      J221...
   .RESULTS...
.RANGEFINDER...
```

Shopping For Data



KitchenAid KHM926CA 9-Speed Digital Hand Mixer with Turbo Beater II Accessories and Pro Whisk - Candy Apple Red
by KitchenAid
387 customer reviews | 79 answered questions

Price: $79.99 Prime

In Stock.
Want it Thursday, Aug. 11? Order within 44 mins and choose Two-Day Shipping at checkout. Details
Ships from and sold by Amazon.com. Gift-wrap available.

Color: Candy Apple Red

$79.99 | $79.99 | $79.99 | $79.99 | $79.99

□ Include 3-Year Protection for $6.9
□ Include 2-Year Protection for $5.0

- 9 Speeds: Combine ingredients at speed one for slowly stirring in chunky ingredients, speed six for mixing bread batter and creaming butter and sugar, then speed nine for whipping meringue.
- Soft StartTM Feature brings the beaters up to the selected speed gradually to help prevent ingredients from splattering.
- Lock the cord into either the left or right side of the mixer so you can approach ingredients from any angle. The round cord is easy to wipe clean.
- Soft Grip Handle provides comfort while mixing.
- Easily and quickly remove any accessory from the hand mixer with the single-handed press of a button.

12 new 2 used from $67.99 1 refurbished from $39.99

Ship to:
Joshua Sti- ARLI

Other Sellers on Am

$89.95
+ Free Shipping
Sold by: GUTSY

$99.00
+ Free Shipping
Sold by: DataVision Computer Video

$99.00
+ Free Shipping
Sold by: SANTECH U

15 used & new

Frequently Bought Together

Total price: $113.86

☑ This item: KitchenAid KHM926CA 9-Speed Digital Hand Mixer with Turbo Beater II Accessories and Pro Whisk... $79.99
☑ Chef Essential Stainless Steel Non-Slip Mixing Bowls Set with Handles and Pour Spouts, Set of 3, Red $24.99
☑ Wilton 570-1121 Easy Flex 3-Piece Silicone Spatula Set, Blue $8.88

Customers Who Bought This Item Also Bought

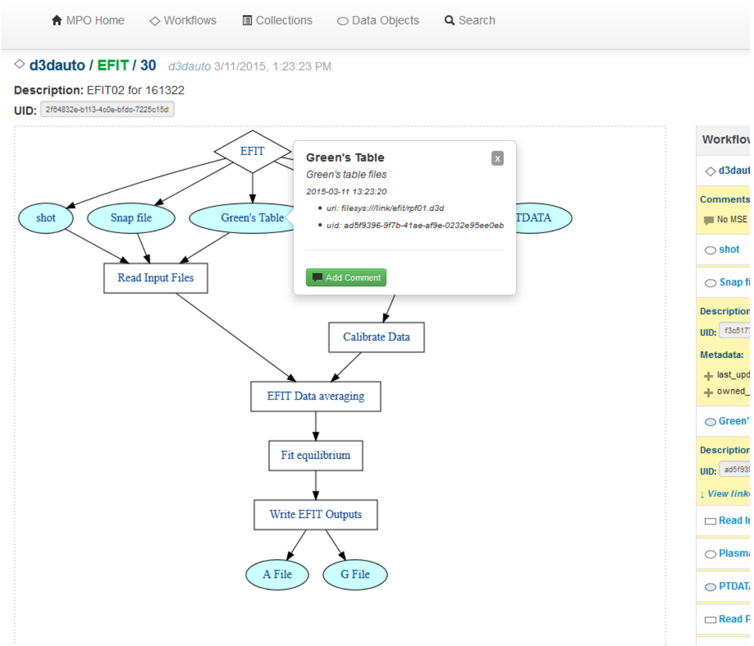
Page 1 of 10

- Online services are very good at helping customers find things they are interested in.
- Search filter and browse
  - Search across multiple criteria
  - Filter by constraint
- Browse ‘related items’
  - Customers who bought this also bought…
  - Customers who looked at this ended up purchasing ...
  - Product reviews

Why not ‘shop’ for interesting useful data ?

Data Relationships are Graphs

- MPO - Metadata Provenance Ontology
- Data provenance represented as directed acyclic graphs



d3dauto / EFIT / 30 d3dauto 3/11/2015, 1:23:23 PM

Description: EFIT02 for 161322

UID: 2f64832e-b113-4c0e-bfdc-7225c15d

Green's Table

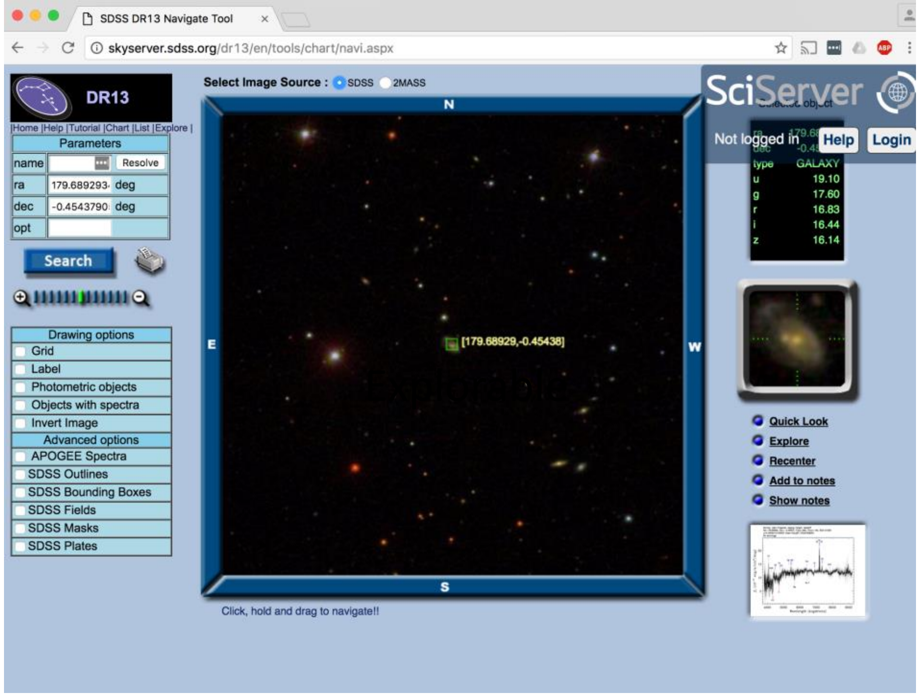
Green's table files

2015-03-11 13:23:20

- *uri: filesys:///link/efit/rpf01.d3d*
- *uid: ad5f9396-9f7b-41ae-af9e-0232e95ee0eb*

Data Relationships are Graphs

Sloan Digital Sky Survey



SDSS DR13 Navigate Tool
skyserver.sdss.org/dr13/en/tools/chart/navi.aspx
DR13
|Home |Help |Tutorial |Chart |List |Explore |
Parameters
Resolve
Search
Select Image Source : ● SDSS ○ 2MASS
N
E
W
S
[179.68929,-0.45438]
Click, hold and drag to navigate!!
SciServer
Not logged in
Help
Login
Quick Look
Explore
Recenter
Add to notes
Show notes

name		
ra	179.689293	deg
dec	-0.4543790	deg
opt		

Drawing options
- □ Grid
- □ Label
- □ Photometric objects
- □ Objects with spectra
- □ Invert Image

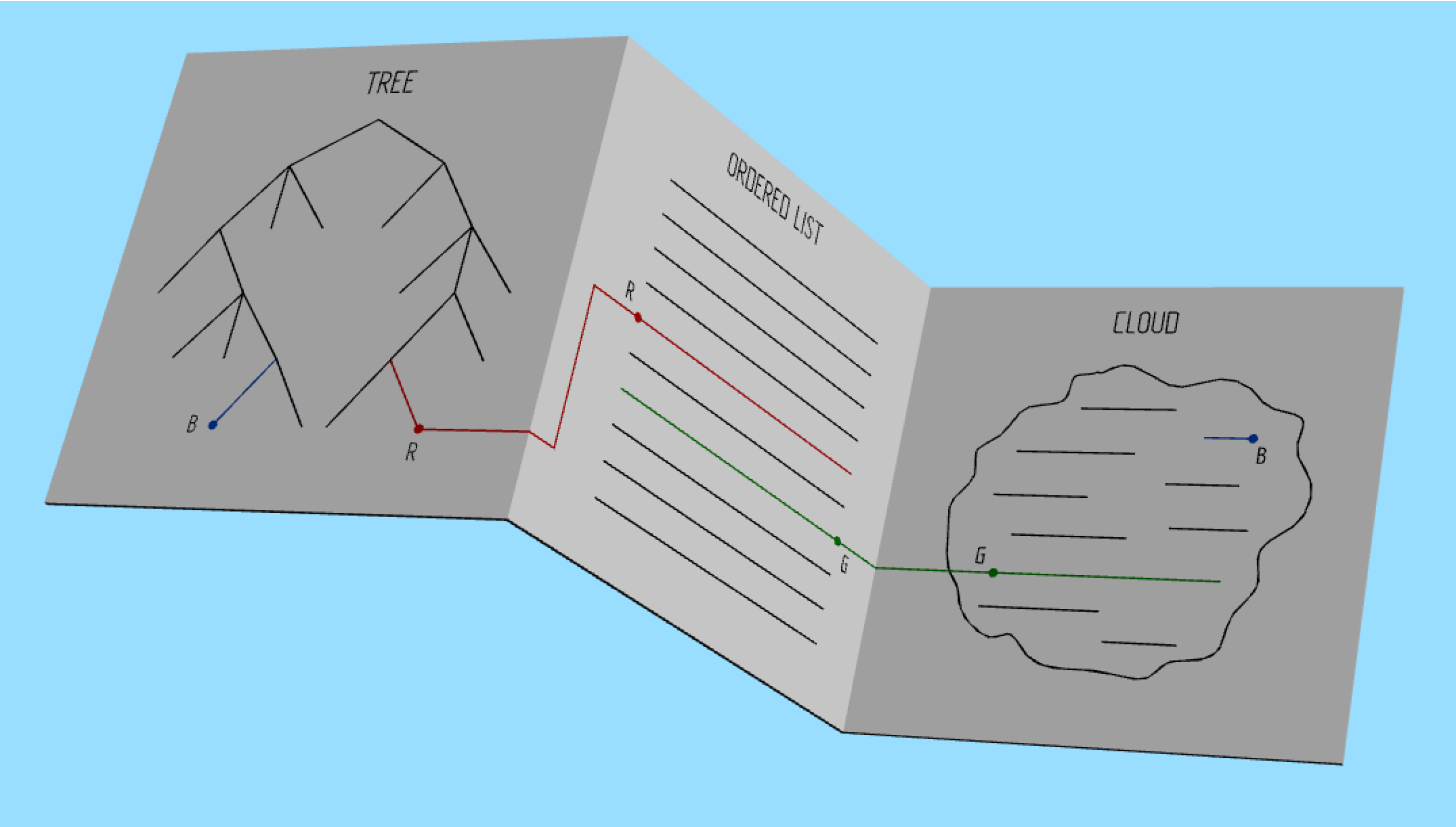
Advanced options
- □ APOGEE Spectra
- □ SDSS Outlines
- □ SDSS Bounding Boxes
- □ SDSS Fields
- □ SDSS Masks
- □ SDSS Plates

Selected object

ra	179.68
dec	-0.45
type	GALAXY
u	19.10
g	17.60
r	16.83
i	16.44
z	16.14

These implementations tend to be purpose built.

Navigational Metadata



Search and navigate within and among different data organizations.

Generalize Data Relationship Tools

- Store schema information – the collection of relationships – as data
  - Provide an API and a GUI populate and explore the data relationship schemas.
- Store instance information – the actual relationships between specific records – as data
  - Provide an API and GUI to populate and explore the data relationship instances.
- Represent all data instances as URIs so that the relationship graphs are agnostic to the type of data being related

Data Granularity

- MPO pointed out a problem with granularity.
  - To compute useful things from the provenance graphs, URIs need to be very specific.
  - To display something interesting/understandable we need to summarize.
- This need to display reduced detail exists in many contexts.
  - Zooming in and out on complex graphs, timelines and maps.
  - Zooming in and out on maps

Sharing Tools Within and Between Communities

- Within a community, scientists can easily make sense of the data at all research facilities in their field of expertise.
  - They know how to use the tools
  - The tools provide data connections needed to understand results
- Developers can share efforts and entire community can take advantage of them.
- MDSplus has been both of these things in the magnetic fusion research community.
- To realize the former, it is likely a community would adopt shared schemas to describe their experiments. Users would then know how to work at other sites, with other data sets.
- The latter will enable disparate research communities (Fusion Energy, Earth Science, Social Science) to leverage each other’s conceptual and implementation development work.

Costs and Mitigations

- For these metadata to be useful and interesting, they have to be populated.
  - This will take effort on the part of the users.
  - The benefits of that effort will not be realized until the metadata exists.
  - The primary beneficiaries of this will probably not be the people doing this work.
  - They will each benefit from each other’s efforts.
- An easy entry slope
  - Encode existing data relationship systems
  - Mine them for initial data sets
  - Populate them automatically where possible

We are Just Getting Started

The project is funded by NSF starting Oct ’16

- What are we missing?
- What is not going to work?
- What other kinds of data relationships should we support?
- Are there data that can not be described by URIs?
- Your questions ?