Science Organizer

A Semantics-based Information Repository for Distributed Scientific Investigations

Richard M. Keller, Ph.D.
Principal Investigator
richard.keller@nasa.gov

Semantic Technologies Applied Research Team
Intelligent Systems Division
NASA Ames Research Center

Funded by NASA R&D Programs and the NASA Astrobiology Institute

http://sciencedesk.arc.nasa.gov/scidesk/

Contributors

Core Team

- Dan Berrios
- Robert Carvalho
- David Hall
- Rich Keller
- Ian Sturken
- Shawn Wolfe

Former Teammates

- Jim Chen
- Mike Compton
- Jon Guice
- Dennis Heher
- Kim Hubbard
- Larry Kiser
- Deepak Kulkarni
- David Nishikawa
- Steve Rich
- Keith Swanson
- Sergey Yentus

Key Science Collaborators

- Brad Bebout (ARC)
- Dave Blake (ARC)
- Dave Des Marais (ARC)
- David McKay (JSC)

POC: Dr. Rich Keller, richard.keller@nasa.gov

Problem:
Capturing the Scientific Record in Distributed Teams

Science Organizer

Accurate capture and preservation of the scientific record during an investigation is critical to conducting research, replicating results, & conducting follow-on studies, yet is difficult to achieve:

- Distributed teams
- Wide variety of data sources (sensors, instruments, data analysis software)
- Heterogeneous information formats (image, data)
- Distributed data curation
- Lack of metadata
- Different terminology used by different groups
- Difficult to see the connections between data/information gathered by different people

POC: Dr. Rich Keller, richard.keller@nasa.gov

Semantic Technology Solution

1. Centralize: Store all project information in a common repository *(at least virtually)*
2. Standardize: For each project team, engineer a common information model and vocabulary for expressing and interrelating all project information – not focusing solely on data
3. Contextualize: Annotate the data with ample semantic context to provide a basis for understanding interrelationships

Develop a Semantically-structured Information Repository: ScienceOrganizer

POC: Dr. Rich Keller, richard.keller@nasa.gov

What is ScienceOrganizer?

- **A Web-based collaborative knowledge management tool for distributed teams of scientific investigators**
- **Facilitates information sharing, integration, correlation**
- **Serves as a project information repository / digital library:**
  users upload/download heterogeneous project information products – images, datasets, documents, and various types of scientific records (describing samples, field sites, measurements, instruments, etc.)
- **Features contextualized cross-linkage:**
  enables rapid access to interrelated information; *permits linking data and observations to scientific hypotheses*
- **Supports inference capabilities**:
  permits formal reasoning about the repository contents
- **Functions as a project archive:**
  tracks history of project team's fieldwork, labwork, and associated data collection activities

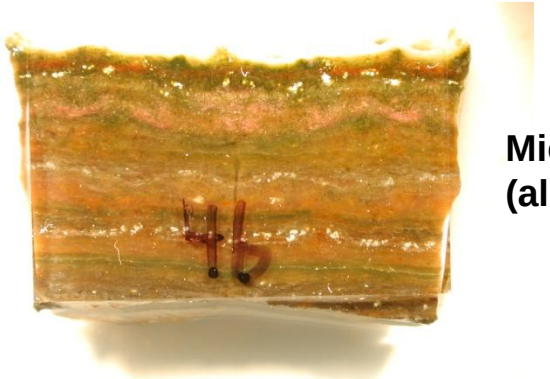
POC: Dr. Rich Keller, richard.keller@nasa.gov

Scientific Activities Supported using ScienceOrganizer

- Collecting scientific field data
- Performing laboratory experiments
- Acquiring data from remote instruments
- Archiving samples and images
- Tracking scientific hypotheses
- Writing scientific papers & proposals
- Conducting education and outreach

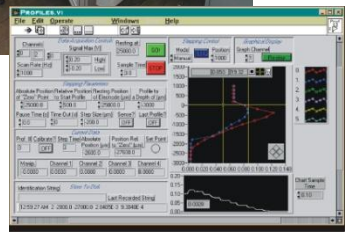
POC: Dr. Rich Keller, richard.keller@nasa.gov

Example:
Early Microbial Ecosystems Investigation



Microbial mat
(algae)

Collection of
microbial mats
in the field



Detailed studies of
mat biogeochemistry

- monitoring
- analysis
- experimentation

Greenhouse Incubator

Trace gas production and
consumption under
“Early Earth” conditions

geographically-disbursed
team of collaborators

POC: Dr. Rich Keller, richard.keller@nasa.gov

Preserving the Evidence

- **Mat “4b” is a piece of evidence in the scientific investigation**
- **Want to capture all contextual information about Mat “4b” for use in:**
  - **conducting analyses**
  - **generating hypotheses**
  - **conveying situational information to remote collaborators**
  - **writing papers**
  - **designing follow-on field investigations**

POC: Dr. Rich Keller, richard.keller@nasa.gov

What is known about Mat "4b"?

Science Organizer

Spring Beach

Brad Bebout

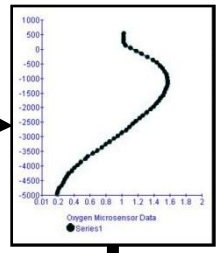
Greenhouse

collected-by

collected-at

stored-in

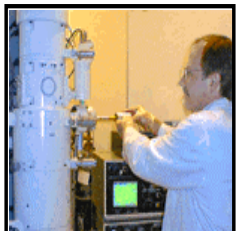
O_2 Concentration



has-measurement

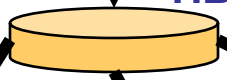
has-culture

measured-with



Electron Microscope

HBC-2 Microbial culture

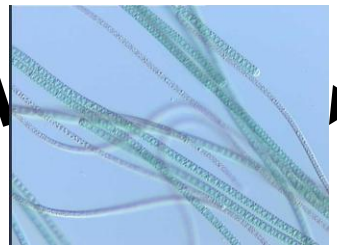


imaged-with

has-image

cultivated-by

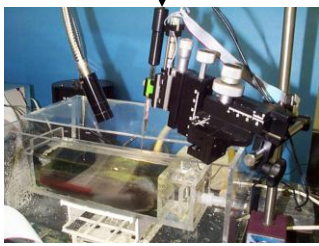
has-recipe



Mary Hogan

Culture prep B notes for Lee

Culture recipe



O_2 Microsensor

POC: Dr. Rich Keller, richard.keller@nasa.gov

Key Idea

To facilitate storage, retrieval and comprehension of scientific data, capture the

semantic context

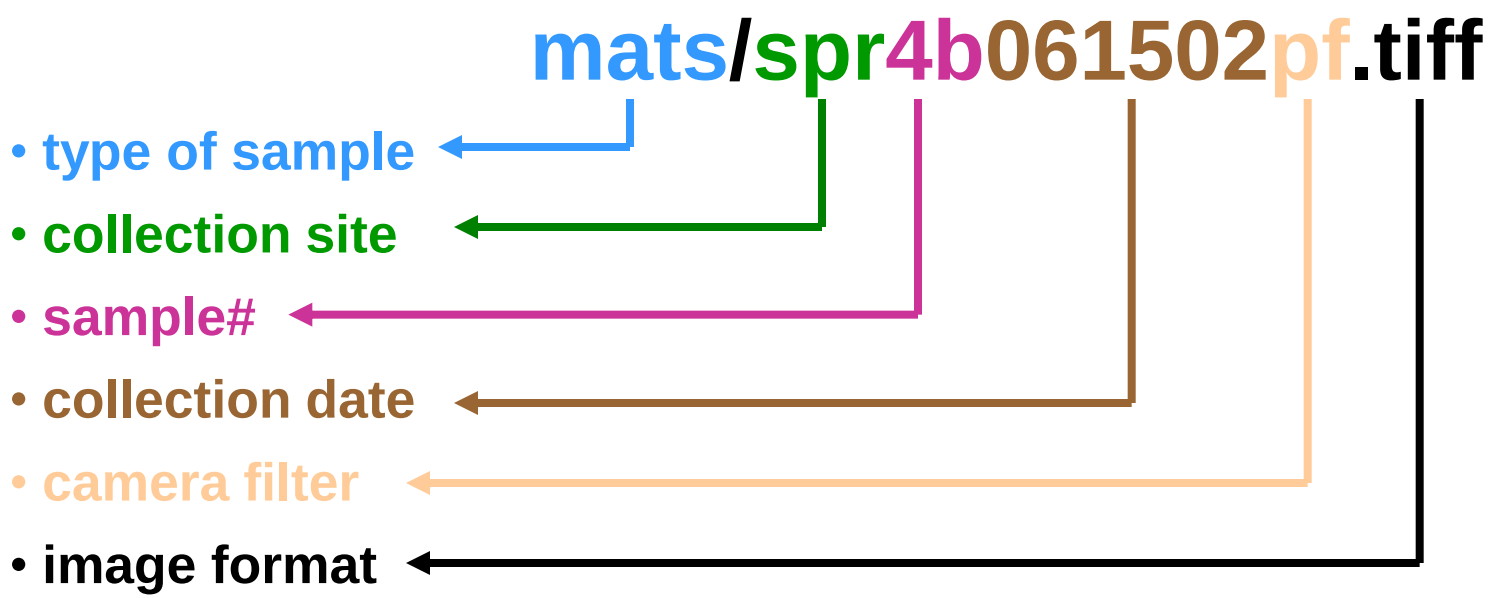
associated with each data product

POC: Dr. Rich Keller, richard.keller@nasa.gov

How is context information typically preserved?

- **Generally, it isn't!**
- **File naming conventions provide some obscure clues:**

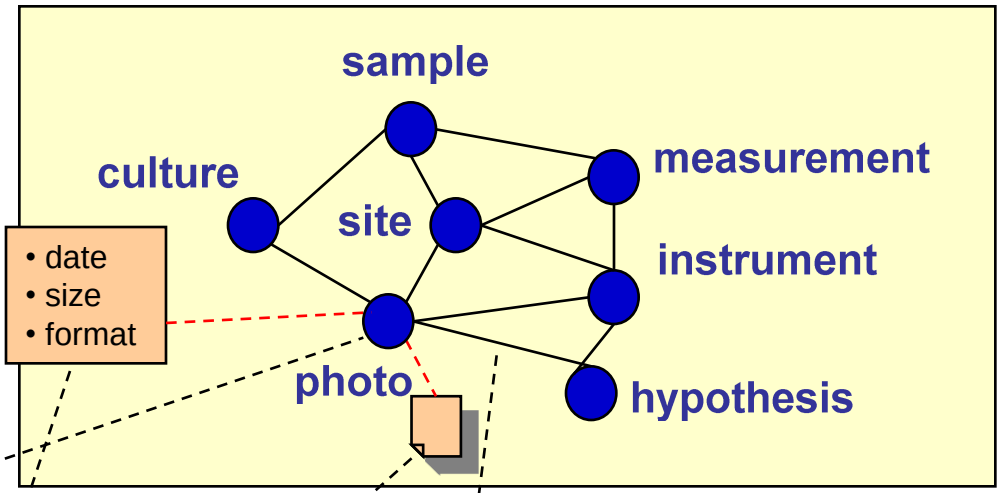
mats/spr4b061502pf.tiff



POC: Dr. Rich Keller, richard.keller@nasa.gov

Semantic Network Structure

Ontology: Specifies the types of nodes, attributes and links defined for scientific investigation

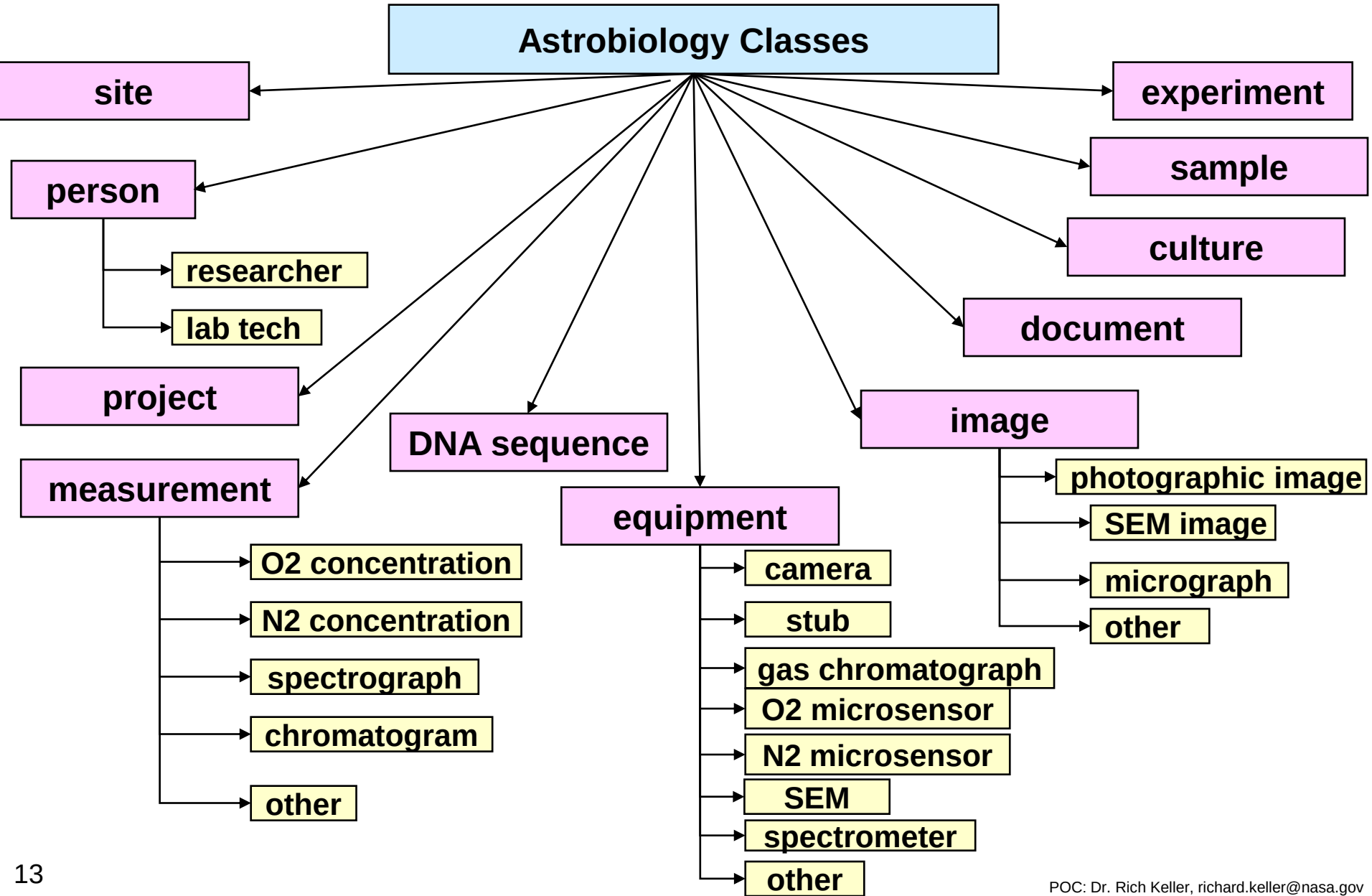


Rules: Add/modify nodes, links & attributes in the network

- **Nodes**: key info resources or organizational structures (describes people, places, measurements, hypotheses)
- **Attributes**: properties of resources (metadata)
- **Links**: relationships among resources (e.g.,"measured by", "supports hypothesis")
- **Attached files**: electronic products associated with resources (in almost any type of format)

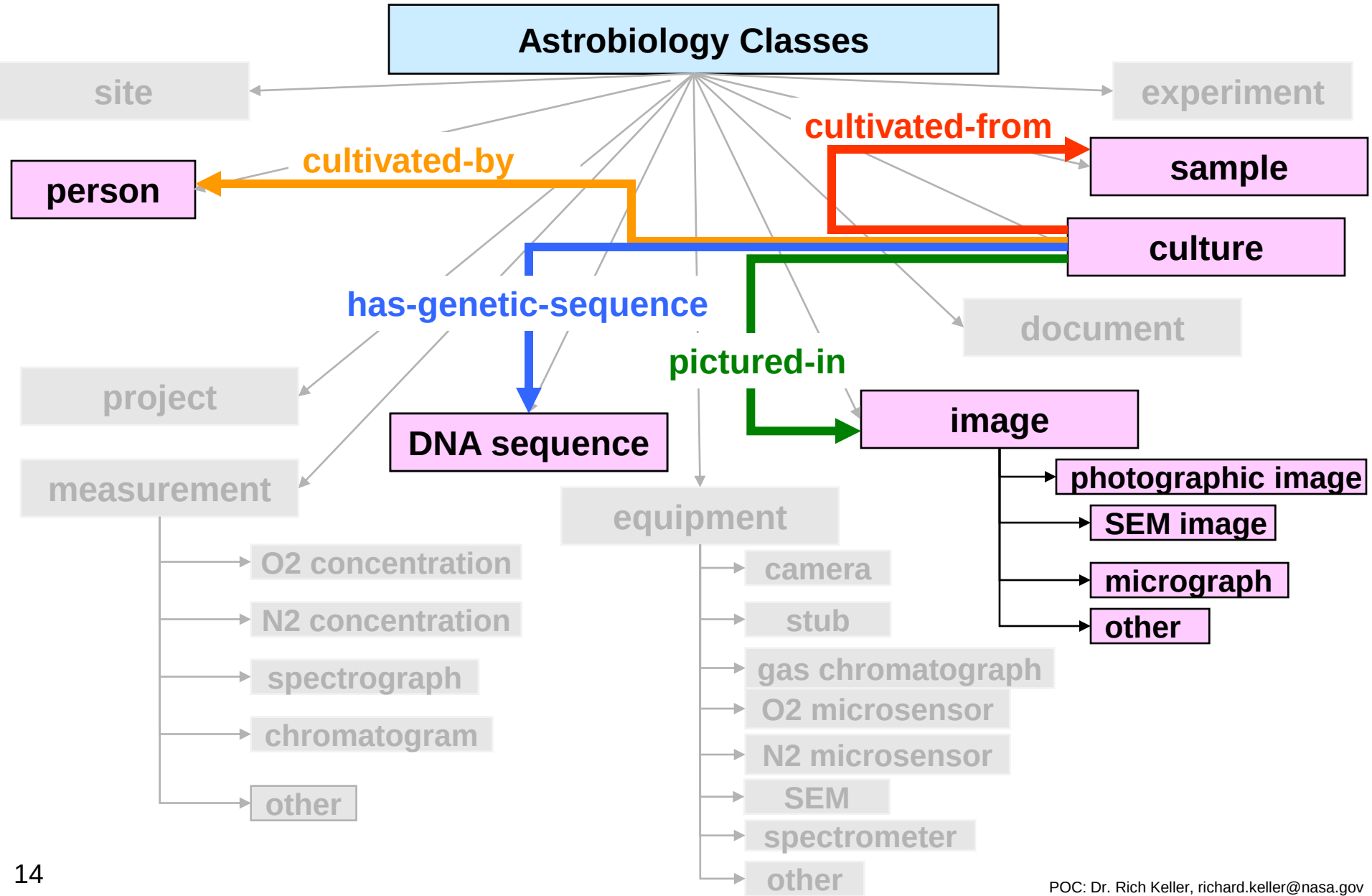
POC: Dr. Rich Keller, richard.keller@nasa.gov

Astrobiology Ontology
(partial)



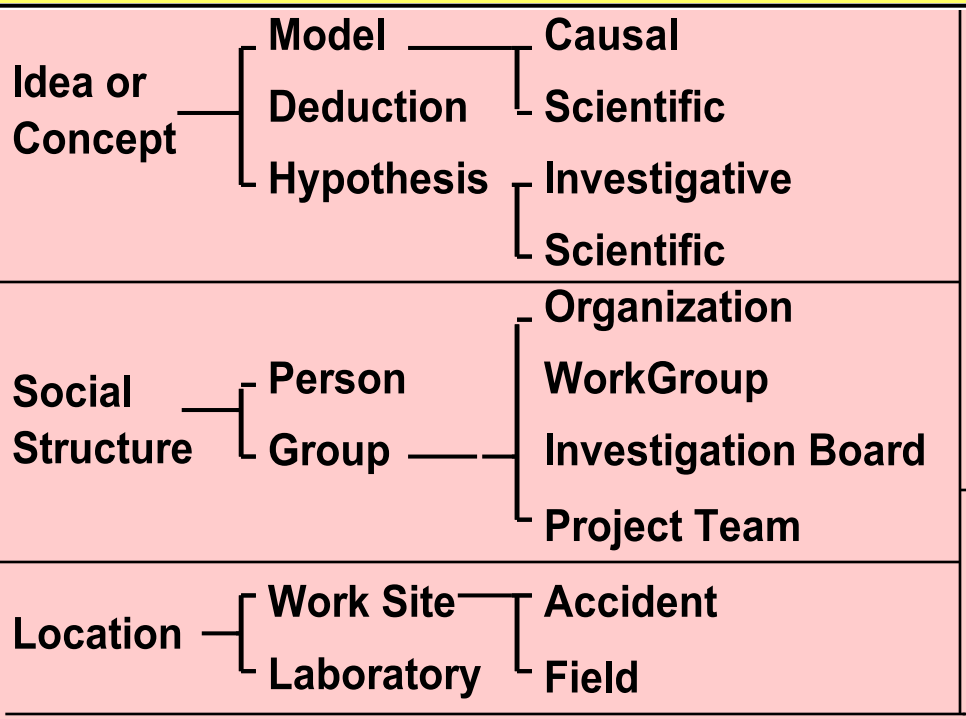
POC: Dr. Rich Keller, richard.keller@nasa.gov

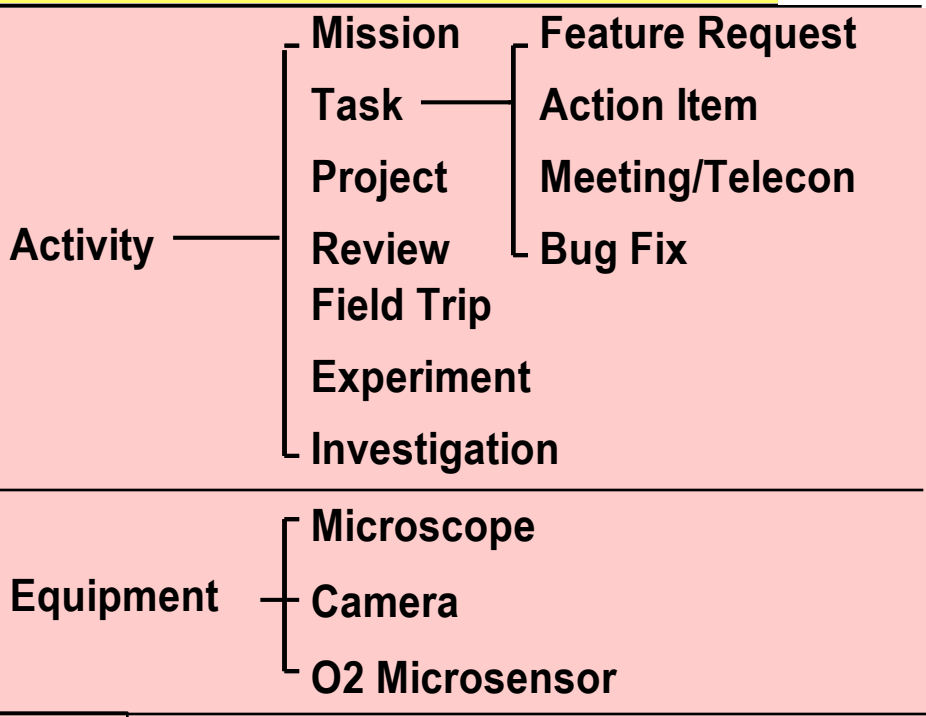
Relationships between Classes

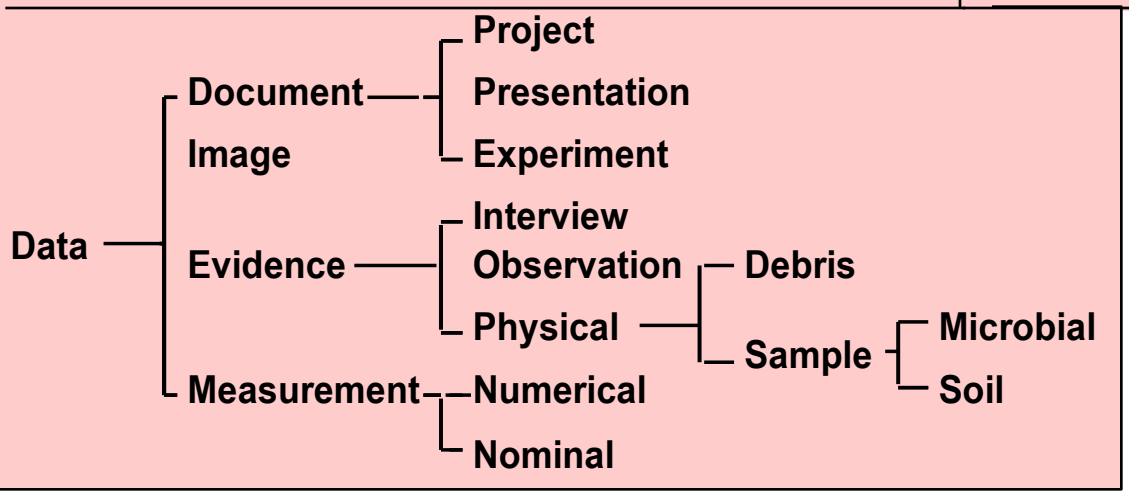


POC: Dr. Rich Keller, richard.keller@nasa.gov

Unified Master Ontology





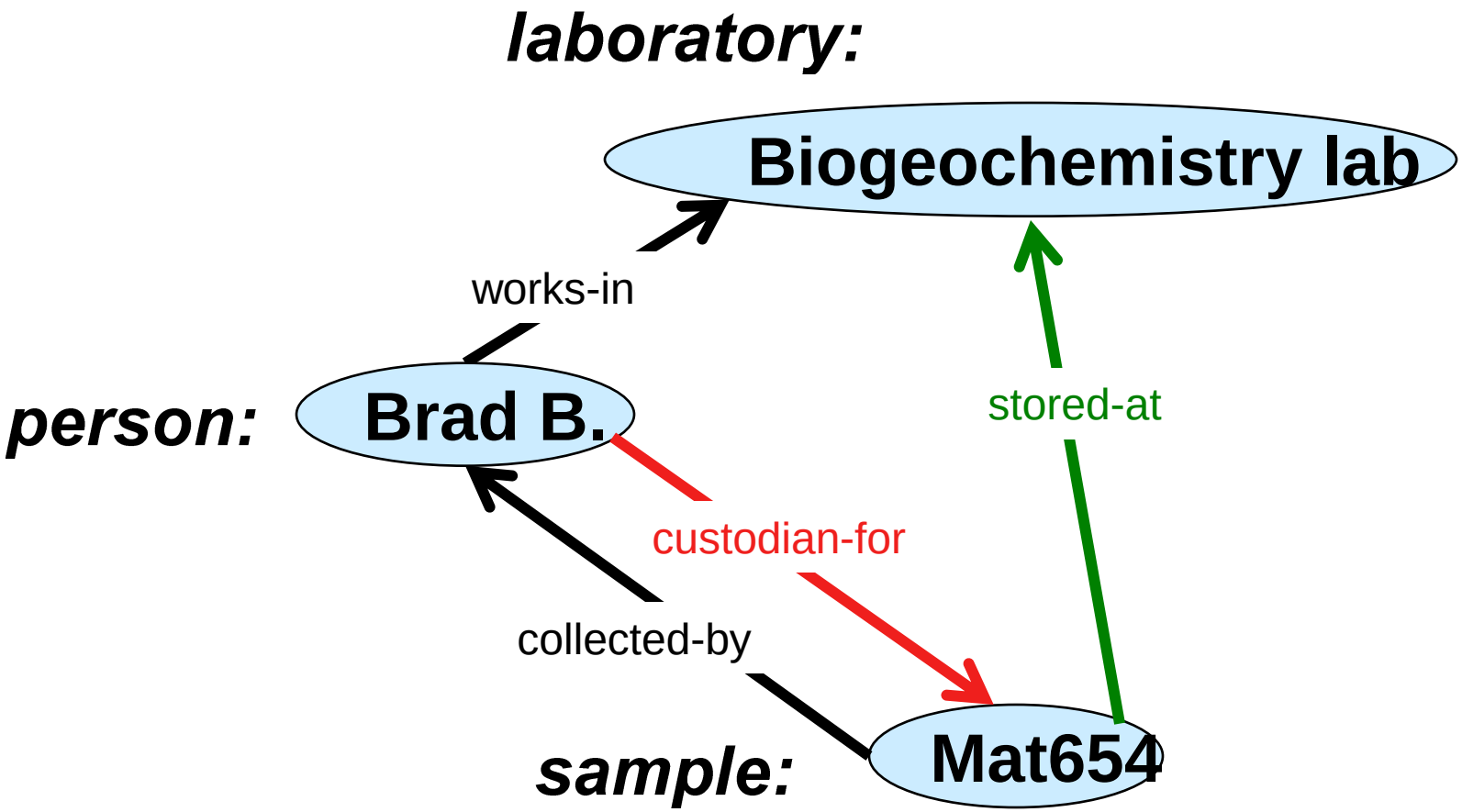


…over 300 classes

…RDFS-like ontology representation language

POC: Dr. Rich Keller, richard.keller@nasa.gov

Simple Example of Rule-based Linking



If **sample S is** ***collected-by*** **person P,** ***and*** **S has no custodian,**
then **P is** ***custodian-for*** **S**

If **person P is** ***custodian-for*** **sample S,** ***and*** **P** ***works-in*** **laboratory L,**
then **S is** ***stored-at*** **L**

POC: Dr. Rich Keller, richard.keller@nasa.gov

SemanticOrganizer's Browser-based Interface (*Algae Mat Sample Node: Spring-M4-b*)

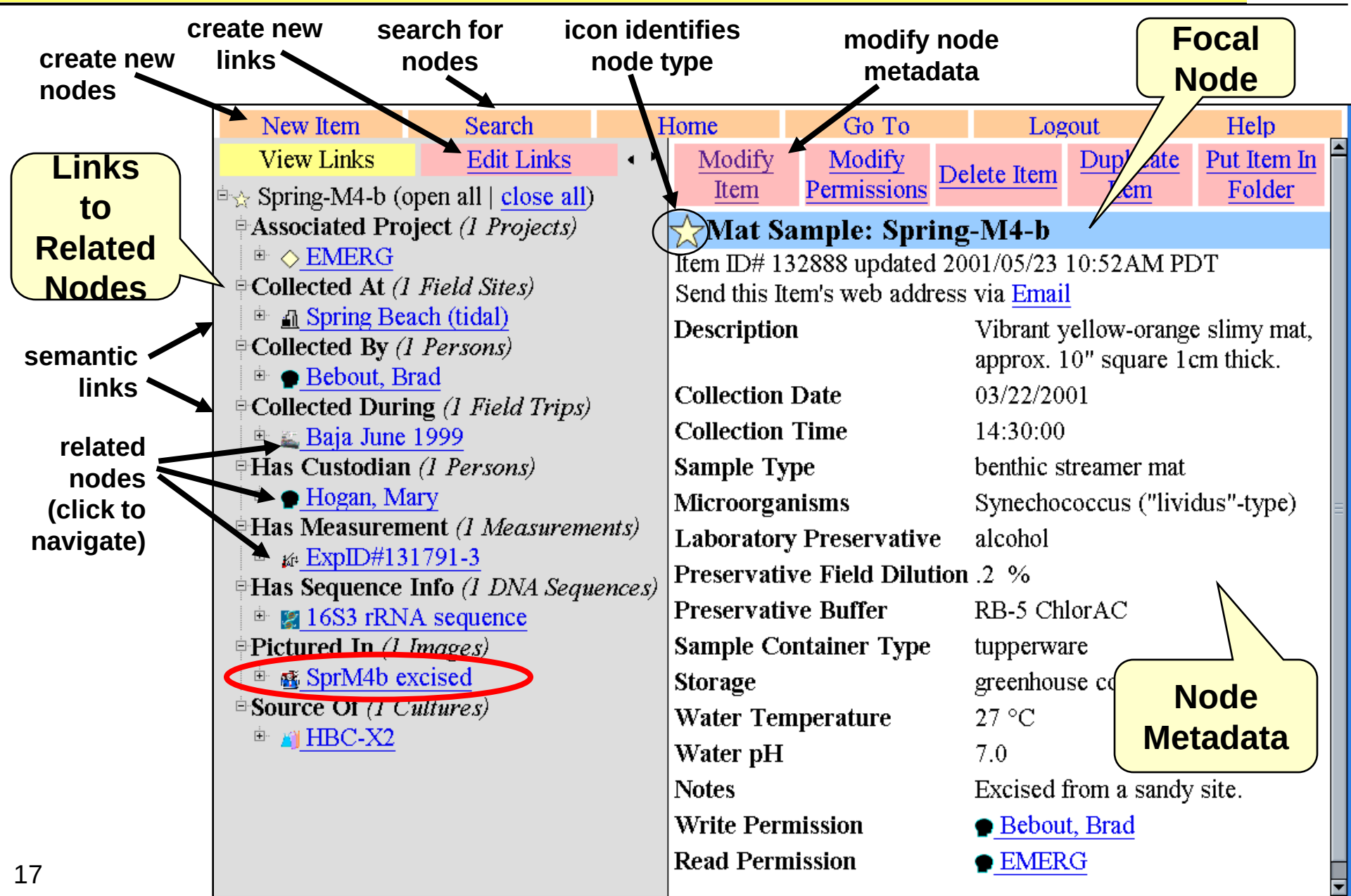


Photo: SprM4b excised

New Item | Search | Home | Go To | Logout | Help

View Links | Edit Links

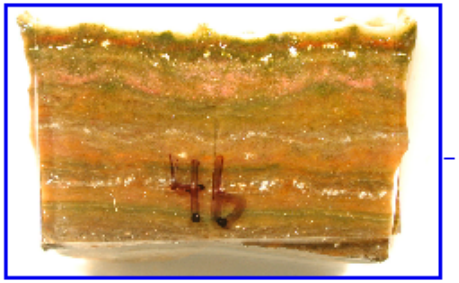
- SprM4b excised (open all | close all)
  - **Camera Used** *(1 Cameras)*
    - Brad's Digital Olympus
  - **Image Of** *(1 Samples)*
    - Spring-M4-b
  - **Imaged By** *(1 Persons)*
    - Bebout, Brad

Modify Item | Modify Permissions | Delete Item | Duplicate Item | Put Item In Folder

Photo: SprM4b excised

Item ID# 132919 updated 2003/06/14 04:28PM PDT

Send this Item's web address via Email



Click image to view at ACTUAL size.

Right Click (or click and hold on Macintosh) to download image.

Click here to download associated jpg file

Help with downloading

Annotate Image

Caption	Sample exhibits classic layering associated with mats from the Baja region.
Image Date	03/22/2001
Notes	Original photo stored in N239 room 433.
Write Permission	Bebout, Brad
Read Permission	EMERG

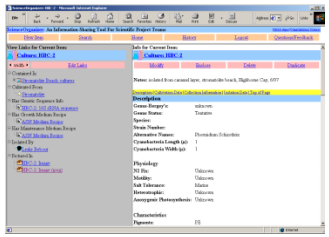
POC: Dr. Rich Keller, richard.keller@nasa.gov

Architecture

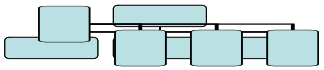
Science Organizer

Interface Layer

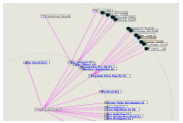
Node Browser Interface



Structure Editors/ Viewers

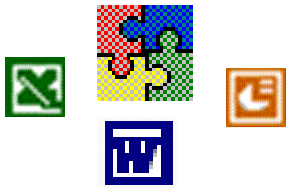


Radial Network Browser



Programmatic/Agent Interface

Microsoft Office Macro Integration



Representation & Reasoning Layer

Email Ingestor
(email & attachments)

Semantic Repository
(items, links, attribute values)

Ontology
(item types, link types, attributes, rules)

Semantic Annotation

WordNet thesaurus

Inference Engine

Implementation Layer

Java servlets

Apache-Tomcat

MySQL Database

Jess

file system

images
documents
datafiles

POC: Dr. Rich Keller, richard.keller@nasa.gov

Some Notable Features

- Text and Semantic search
- Email discussion list integration
- Automated semantic hyperthreading
- Microsoft office integration
- Collaborative image annotation
- Interoperation w/external agents

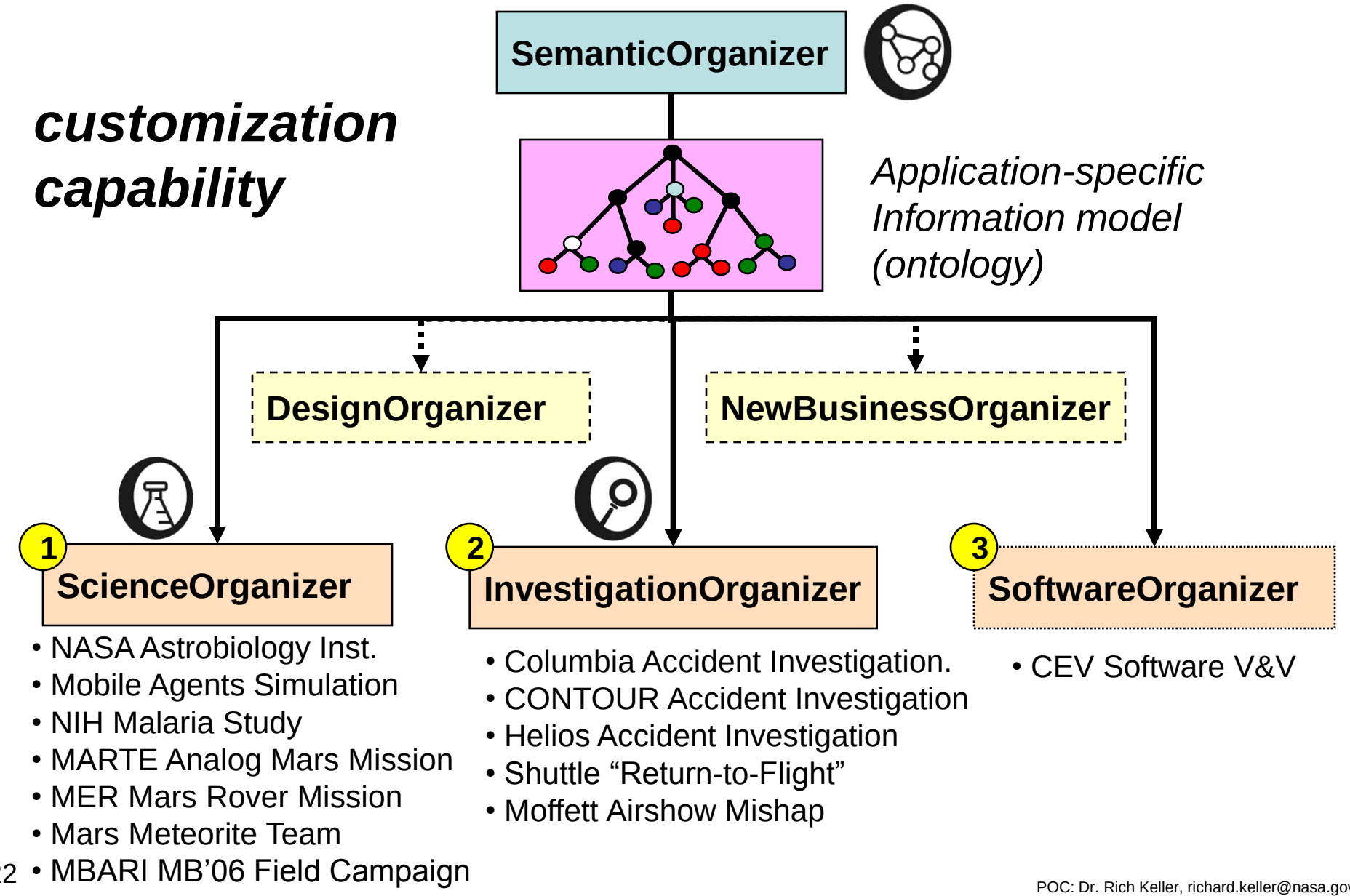
POC: Dr. Rich Keller, richard.keller@nasa.gov

Operational Status

- System deployed initially in 2001
- Over 400 registered individual users from over 50 organizations within NASA at peak
- Over 30 projects hosted
- Over 45,000 nodes & 160,000 links in repository
  - Over 10,000 electronic files stored (documents, image, datasets)
  - Over 10,000 archived email messages

POC: Dr. Rich Keller, richard.keller@nasa.gov

ScienceOrganizer: A SemanticOrganizer Application

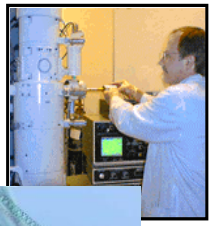


POC: Dr. Rich Keller, richard.keller@nasa.gov

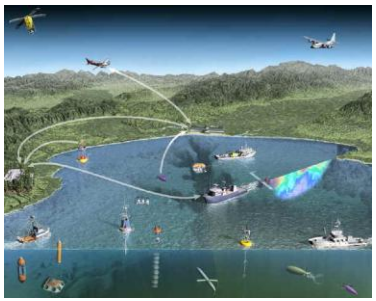
ScienceOrganizer Deployments

Astrobiology Field and Lab Science

NIH-NASA
African Malaria Study



Electron Microscopy Image Archive (Martian Meteorites)



Monterey Bay Oceanographic Campaign

Simulated Mars Surface Exploration

MARTE Mars Analog Drilling Mission

Mars Exploration Rovers Hypothesis Management (prototype)

POC: Dr. Rich Keller, richard.keller@nasa.gov

Questions?

Web Site http://sciencedesk.arc.nasa.gov

Paper

Keller et al.: “SemanticOrganizer: A Customizable Semantic Repository for Distributed NASA Project Teams”, International Semantic Web Conference, Hiroshima, Japan, November 2004.

Honors

ISWC

Winner, Best paper award, 2004 International Semantic Web Conference

Finalist in the 2003 NASA Software of the Year Competition and recipient of a Space Act Award for significant technical contribution to NASA.

Winner of a 2004 NASA Turning Goals Into Reality Award for outstanding accomplishment in supporting the Columbia Accident Investigation Board

Finalist in the 2004 Semantic Web Challenge Competition

POC: Dr. Rich Keller, richard.keller@nasa.gov

Additional Slides

- **Mars Exploration Rover application:**
  Science Hypothesis Organizer

- **ScienceOrganizer Features:**
  1. Text and Semantic search
  2. Email discussion list integration
  3. Automated semantic hyperthreading
  4. Microsoft office integration
  5. Collaborative image annotation
  6. Interoperation w/external agents

POC: Dr. Rich Keller, richard.keller@nasa.gov

MER Science Hypothesis Organizer Objective

Improve MER science return by facilitating the science team's hypothesis formation process

Motivation:

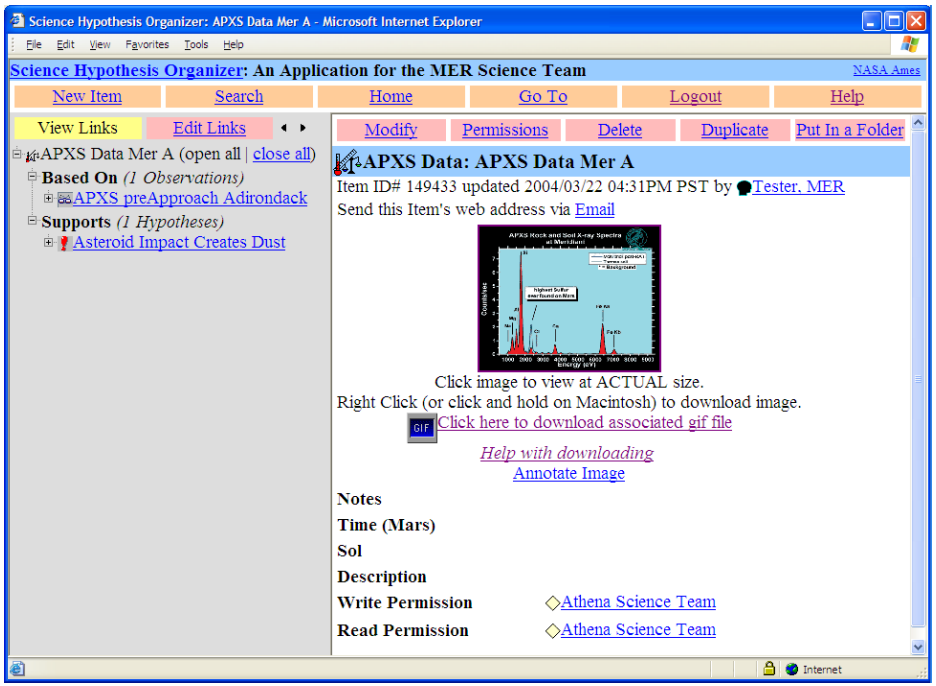
- Hypothesis formation and testing are fundamental drivers for MER science
- Hypothesis formation and reasoning processes are informal and managed in an *ad hoc* fashion
- No representation of the process is formally captured for later analysis or critical reflection
- No centralized catalog of active hypotheses and scientific evidence

POC: Dr. Rich Keller, richard.keller@nasa.gov

MER Science Hypothesis Organizer (SHO)

Permits MER scientists to store/track hypotheses & relate them to:

- observations
- data products
- documents
- & other files stored in mission data systems



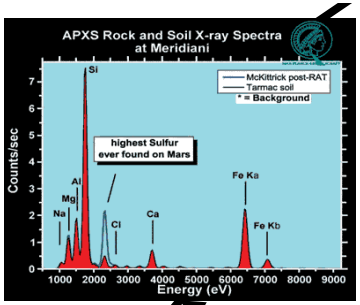
POC: Dr. Rich Keller, richard.keller@nasa.gov

Hypothesis Information Network: *Connecting the dots*

Science Organizer

Hypothesis: Mars surface dust formed from meteor impact

supported-by



APXS data

tested-by

supported-by

proposed-by

John Smith

Adirondack pancam

measured-by

generated-by

supported-by

proposed-by

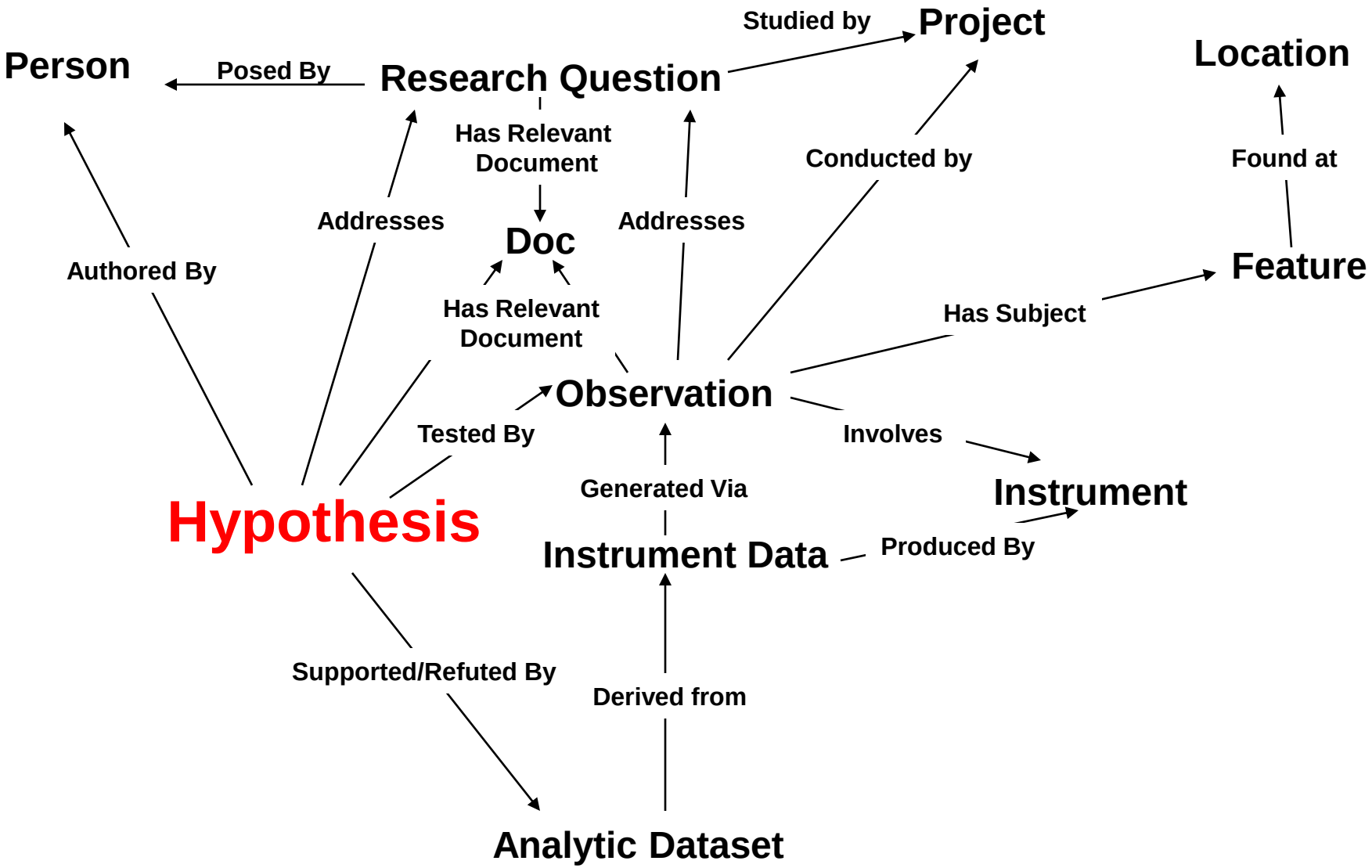
APXS instrument

APXS observation

Hypothesis: Soil is a duracrust of fine sand of aeolian dust & mafic volcanic rock fragments

POC: Dr. Rich Keller, richard.keller@nasa.gov

MER Science Hypothesis Organizer Cross-relationships



POC: Dr. Rich Keller, richard.keller@nasa.gov

MER SHO Object Hierarchy

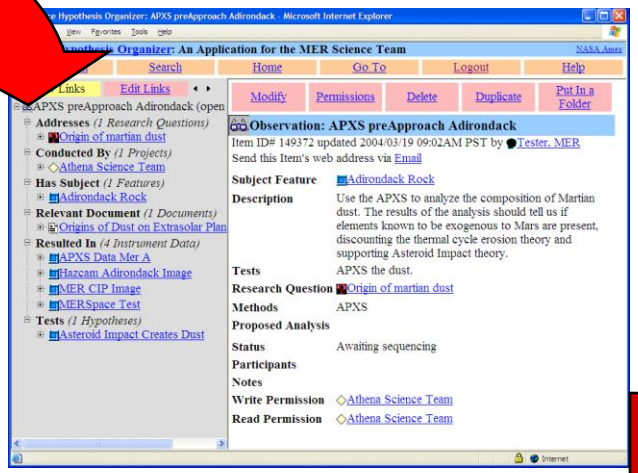
- **All Items**
  - **Analytic Dataset**
  - **Email Message**
  - **Feature**
  - **Hypothesis**
  - **Institution**
  - **Location**
  - **Mailing List**
  - **MER Participant**
  - **Observation**
  - **Project**
  - **Research Question**
  - **Target**
  - **Item (other)**
- **All Instrument Data**
  - **All Images**
    - Hazcam Image
    - Micrograph
    - Navcam Image
    - Pancam Image
    - Image (other)
  - **All Measurements**
    - APXS Data
    - Mini-TES Data
    - Mossbauer Data
    - Measurement (other)
  - Instrument Data (other)
- **All Documents**
  - Meeting Notes
  - Presentation
  - Research Paper
  - Document (other)
- **All Instruments**
  - APXS
  - Camera
  - Mini-TES
  - Mossbauer
  - RAT
  - Rover
  - Instrument (other)

POC: Dr. Rich Keller, richard.keller@nasa.gov

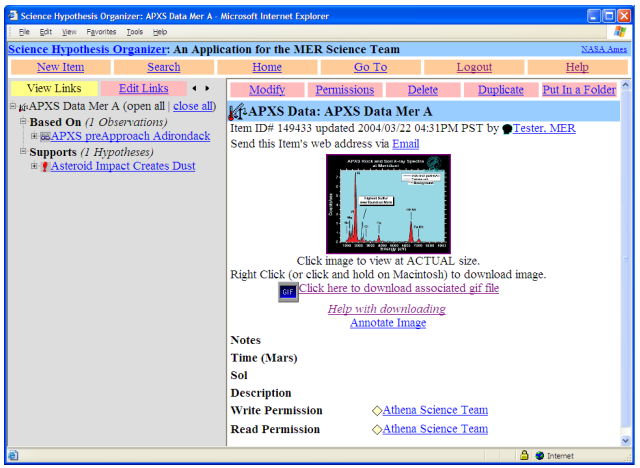
MER SHO Screenshots

Hypothesis

Observation Request

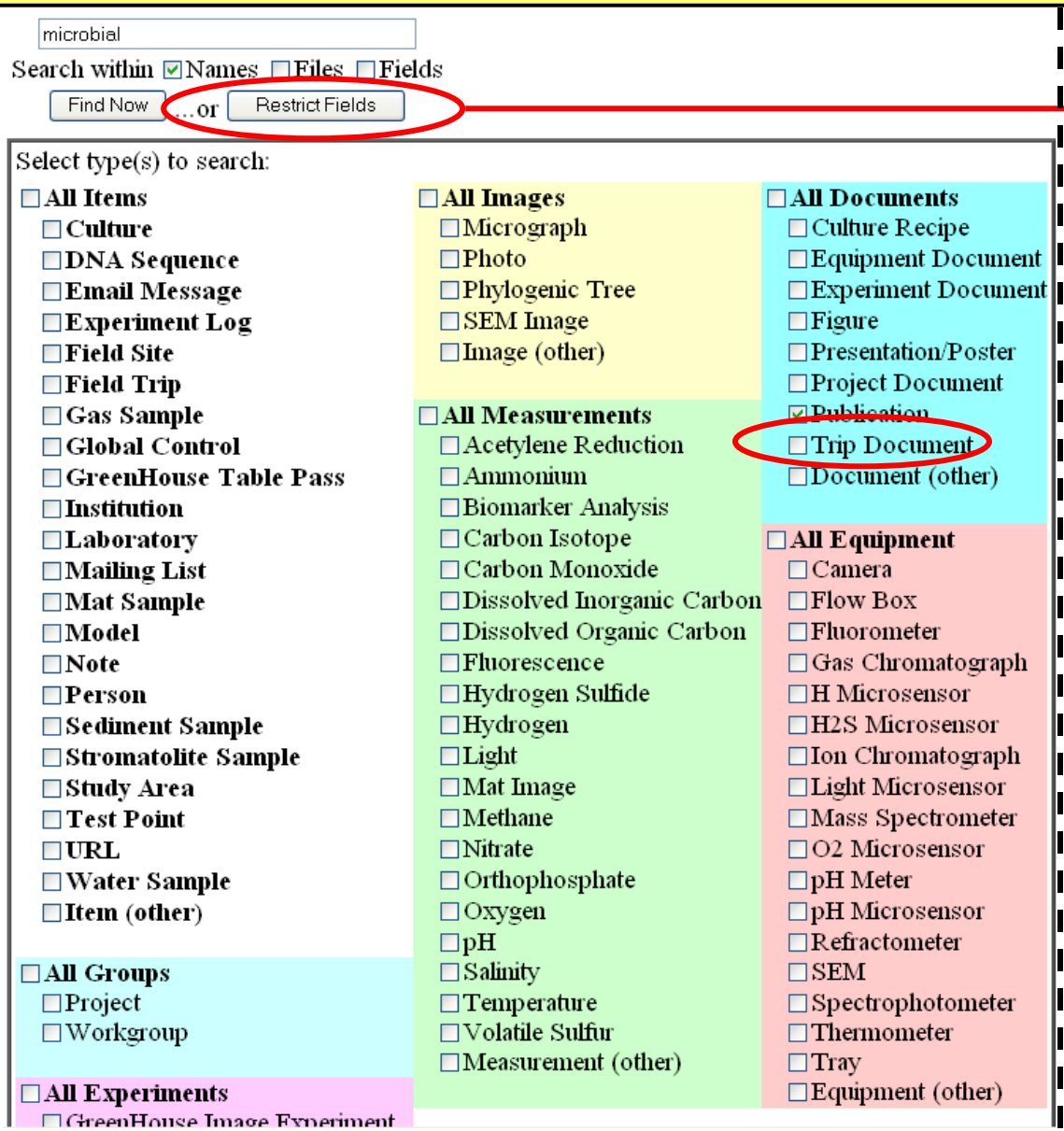


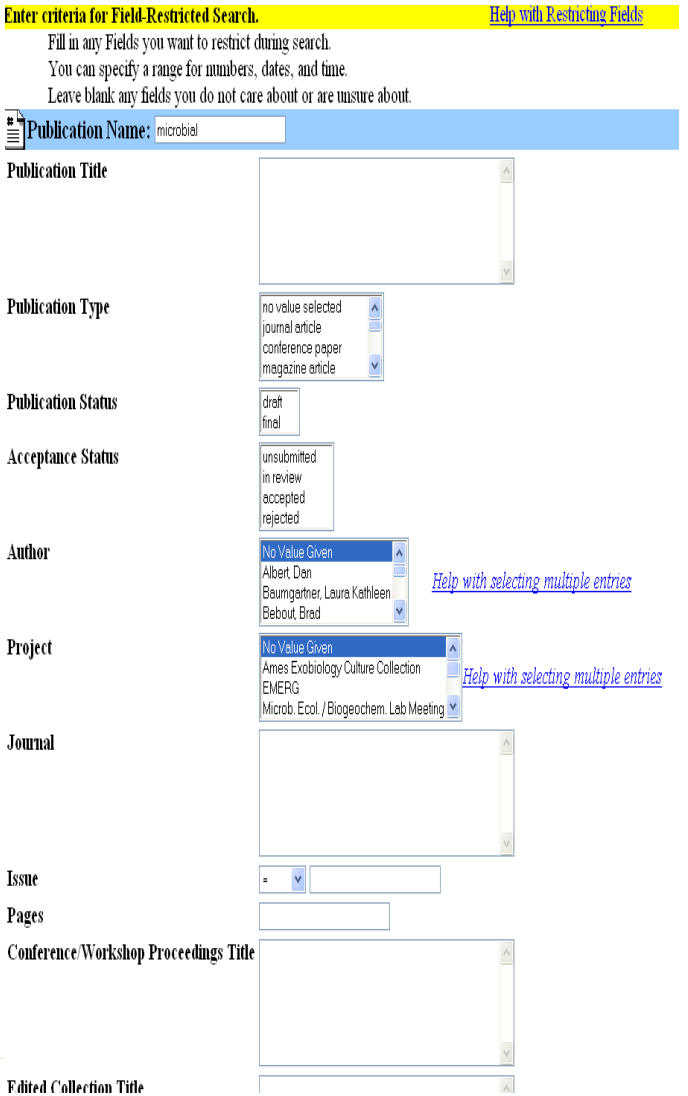
APXS Data Product



POC: Dr. Rich Keller, richard.keller@nasa.gov

Text Search



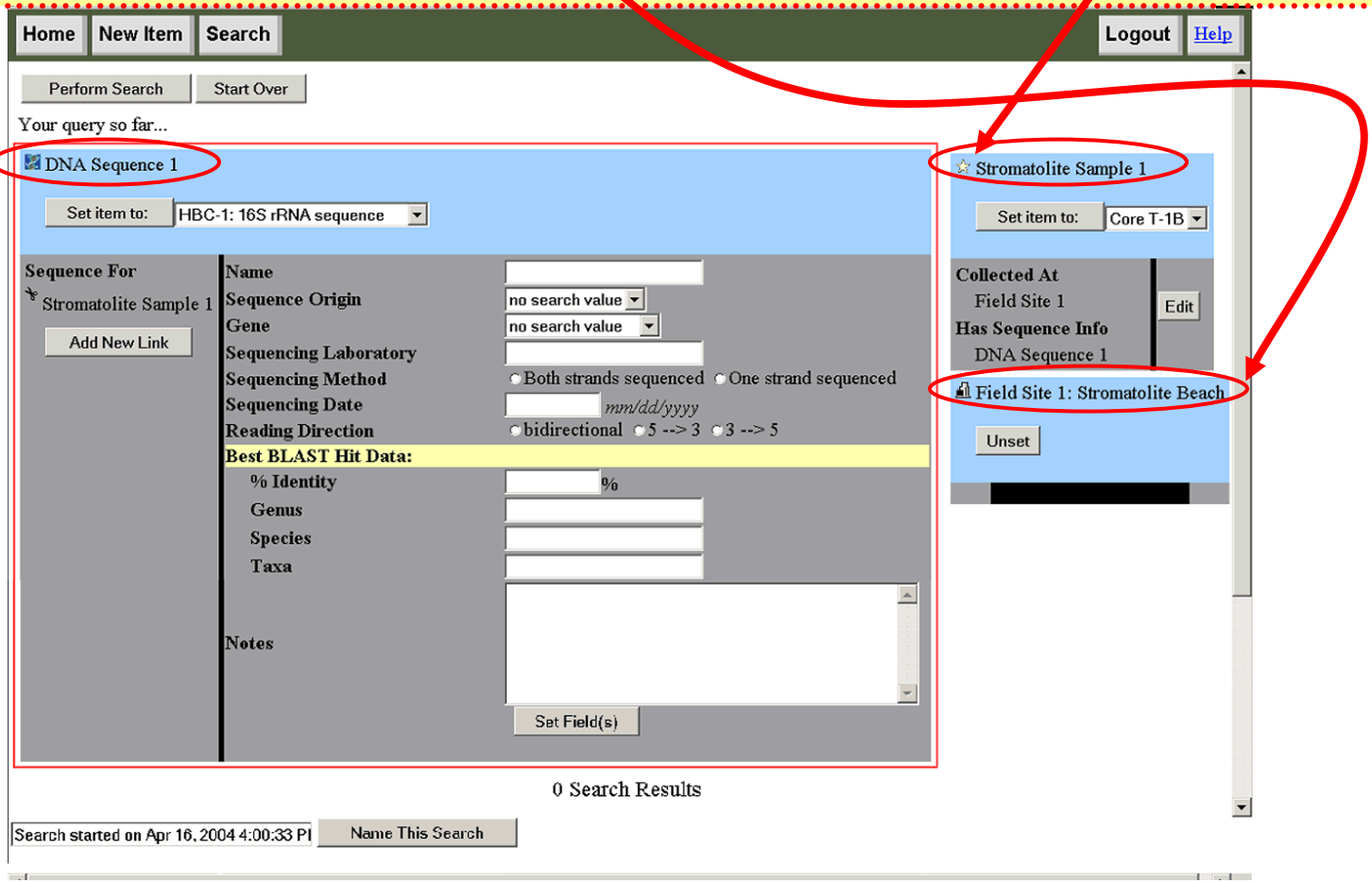


POC: Dr. Rich Keller, richard.keller@nasa.gov

Semantic Search

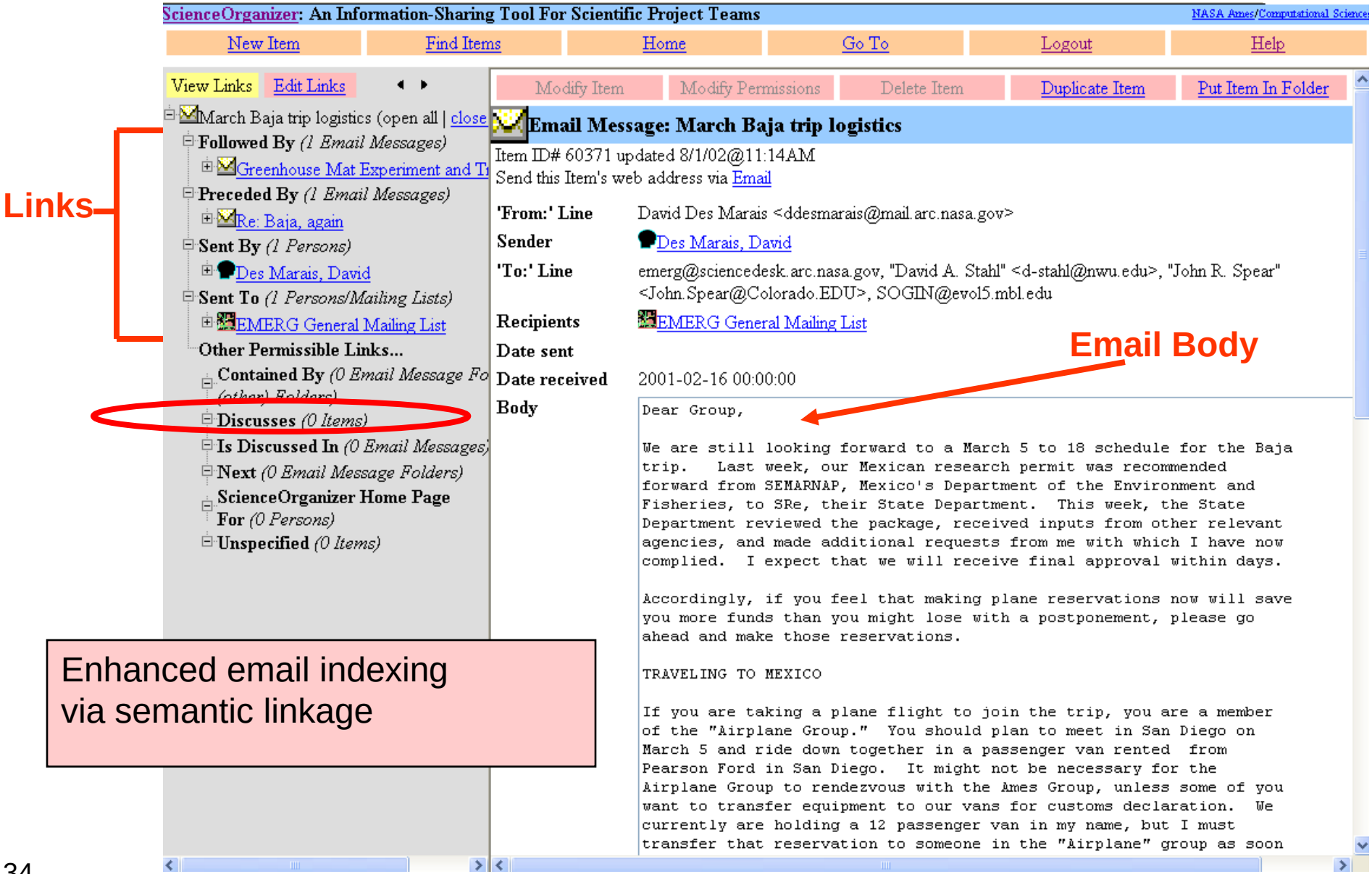
- Search by semantic pattern (patterns involving extended chains of nodes, links, and values)
- Executed by matching pattern against the instance nodes and links in the semantic network

"Find all DNA Sequences associated with Stromatolite Samples collected at the Field Site named "Stromatolite Beach""



POC: Dr. Rich Keller, richard.keller@nasa.gov

Email Discussion List Integration



Enhanced email indexing via semantic linkage

POC: Dr. Rich Keller, richard.keller@nasa.gov

Automated Semantic Hyperthreading [Concept Extraction from Email] (experimental)

Science Organizer

Email body (extract)

"FIELD OBJECTIVES
This field trip will address a number of objectives that include those of the Ecogenomics project as well as those associated with previously-funded grants. Accordingly, we will be performing biogeochemical measurements and collect samples for various analyses back at our respective laboratories…

Straw sampling Protocols

In situ measurements of fluxes:
… number of diel periods : 2 (once in 5 , once in Pond 5 near 6)
… number of chambers: 6 (3 mats, three blanks)
… sampling times: 0600h, 0900h, 1200h, 1500h, 1800h, 2100h, 0000h, 0300h
… Water volume taken per sampling time : 5
… Gas volume taken per sampling interval 5

…SAMPLING
We recognize that the acquisition and pre vation c will be a major objective of this trip. We h good preferences during our November meeting

Relevancy rank	Node	Type
1.063660	Spear, John	(participan)
1.063660	Stahl, David	(participan)
1.063660	Bebout, Brad	(participan)
1.036137	Ames	area)
1.036137	Baja	a)
1.031293	Ecogenomics Focus	(proje
1.020290	Pond 5 near 6	(aque i)
1.020290	Pond 4 near 5	(aque i)
1.009677	Visscher, Pieter	(participan)
1.009677	Hoehler, Tori	(participan)
1.009677	Hogan, Mary	(participan)
1.009677	Garcia-Pichel, Fer	(participan)
1.009677	Dillon, Jesse	(participan)
1.009677	Turk, Kendra	(participan)
1.009677	Miller, Scott	(participan)

Pond 4 near

2. Relevancy ranking score computed:
- **based on heuristic combination of factors**

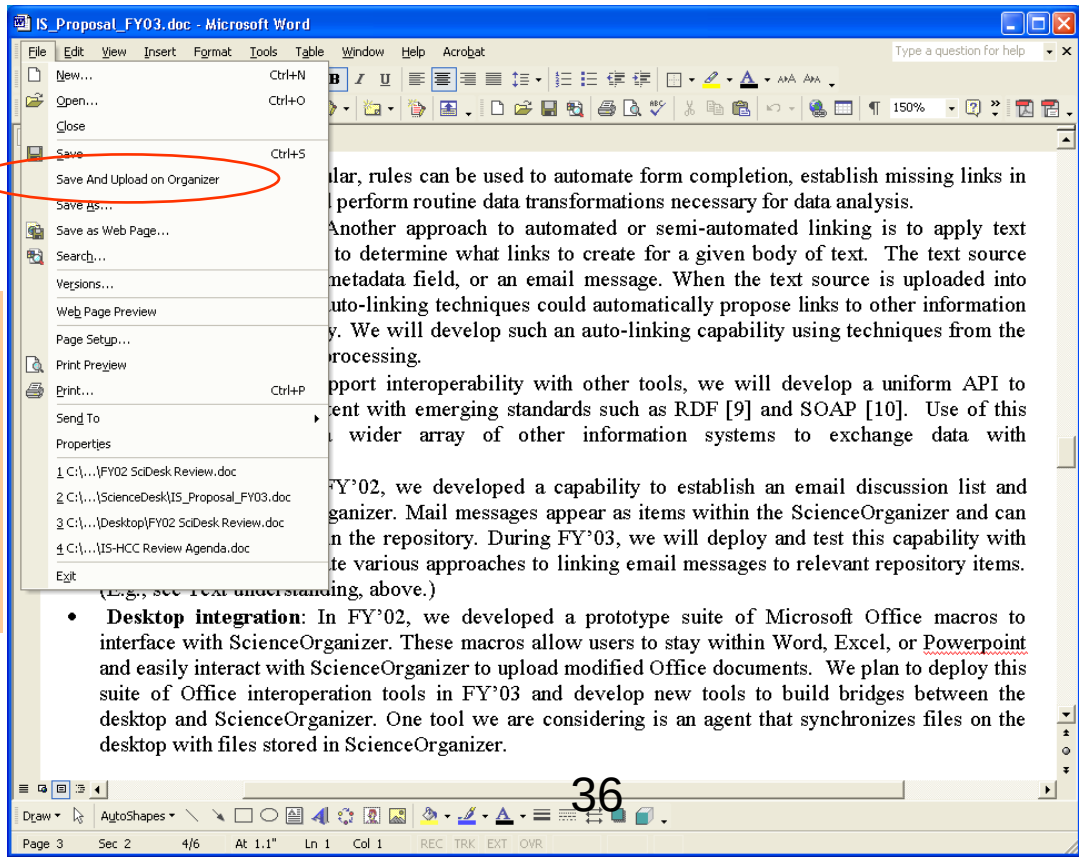
1. Text marked up to identify:
- **nouns matching node names/types**
- **verbs corresponding to links**
- **references to attribute values**
- **WordNet synonyms/hyponyms used to enable inexact semantic matching**

POC: Dr. Rich Keller, richard.keller@nasa.gov

Microsoft Office Interoperation

- **Developed suite of Microsoft Office macros to enable direct upload and subsequent modification of documents without leaving Office application**

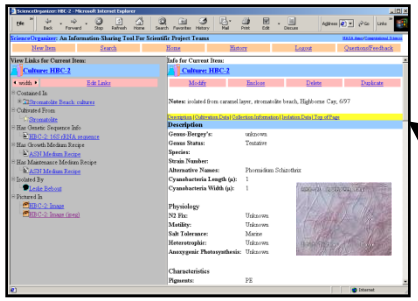
"Save and upload to Organizer"



- Macro communicates w/server
- User fills out metadata using standard Organizer form on creation
- Subsequent saves are transparent

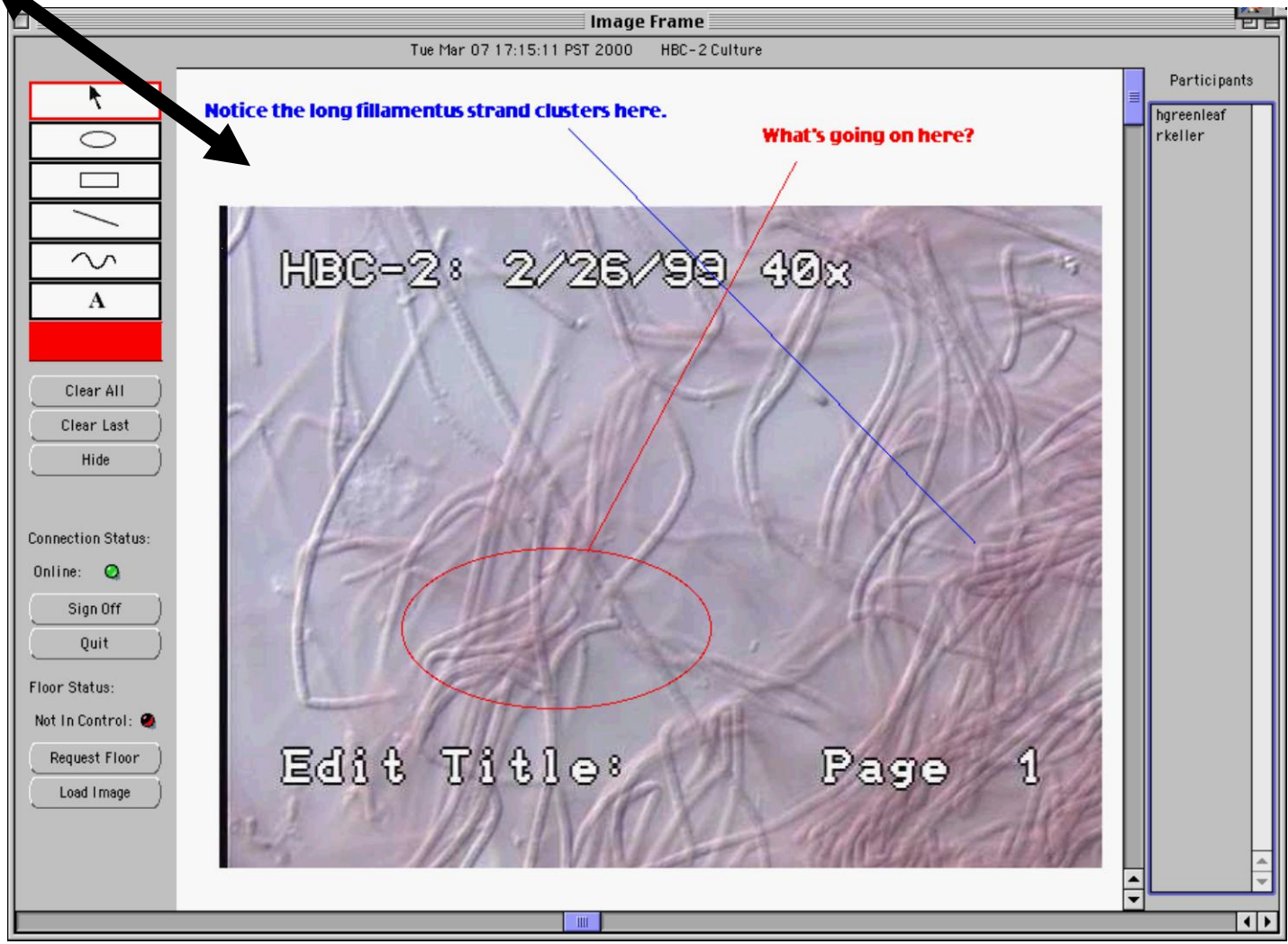
POC: Dr. Rich Keller, richard.keller@nasa.gov

Shared Image Annotator



SemanticOrganizer

Collaborative Whiteboarding & Annotation Tool



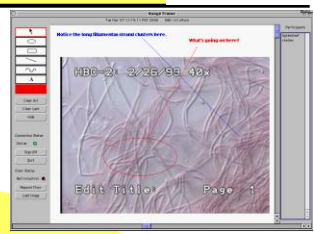
POC: Dr. Rich Keller, richard.keller@nasa.gov

Interoperation w/ External Agents & Systems via API

Science Organizer

Agent-assisted Scientific Experimentation System

Agent-based Mars exploration testbed



images, measurements

image

real-time measurements, images

ScienceOrganizer

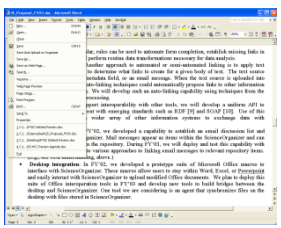
Shared Image Annotator

annotated image

documents

Microsensor Remote Controller

Microsoft Office interface



sensor

POC: Dr. Rich Keller, richard.keller@nasa.gov