

IS-HCC FY'02 Program Review – Oct. 2002

**ScienceOrganizer:
Collaborative Information Management
for Scientific Teams**

Richard M. Keller, Ph.D.

***Information Sharing and Integration Group
Collaborative and Assistant Systems Tech Area
Computational Sciences Division
NASA Ames Research Center***



Outline



- **ScienceOrganizer: brief introduction**
- **FY02 Work:**
 - **Integration with Work Tools & Systems**
 - **Automated Knowledge Acquisition**
 - **New Customers / scale-up**



ScienceDesk Project Staff



Rich Keller, PI

Shawn Wolfe

David Hall

Dan Berrios

Robert Carvalho

Steve Rich

Deepak Kulkarni

Sergey Yentus

Keith Swanson

Ian Sturken

Ling-Jen Chiang

David Nishikawa

Linda Andrews

***Computational Sciences Division
NASA Ames Research Center***

Brad Bebout, Co-I

Steve Carpenter

***Exobiology Branch
NASA Ames Research Center***

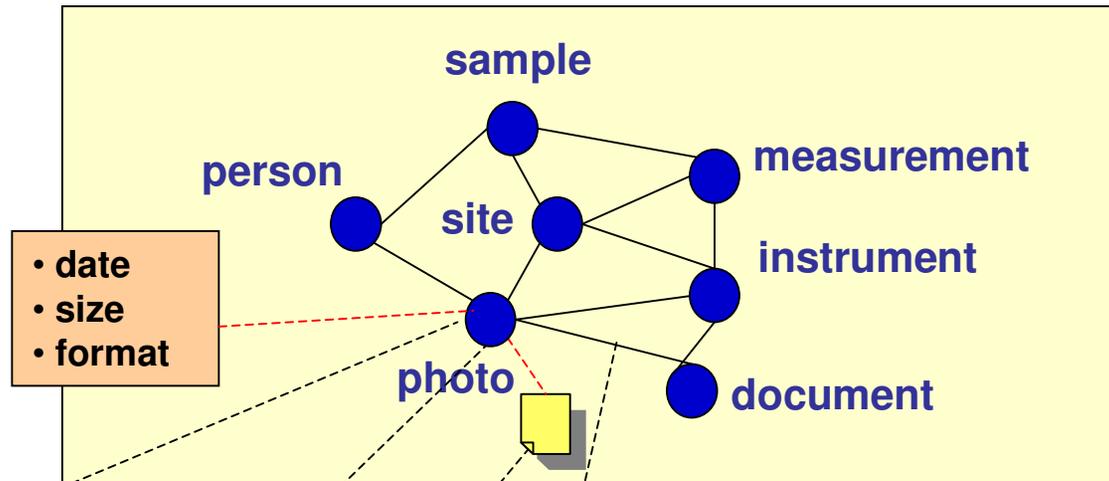


What is ScienceOrganizer?



- **A collaborative knowledge management tool for distributed scientific project teams**
- **A project information repository / digital library:** stores heterogeneous project information products -- images, datasets, documents, and various types of scientific records (describing samples, field sites, measurements, instruments, microbial cultures, etc.)
- **A hybrid tool combining the functionality of:**
 - a database
 - a document-sharing system
 - a hypermedia information space
 - a semantic network } *semantic web-like system*
- **Features cross-linkage:** enables rapid access to interrelated information
- **A “project memory” system:** tracks history of project team’s fieldwork, labwork, and associated data collection activities

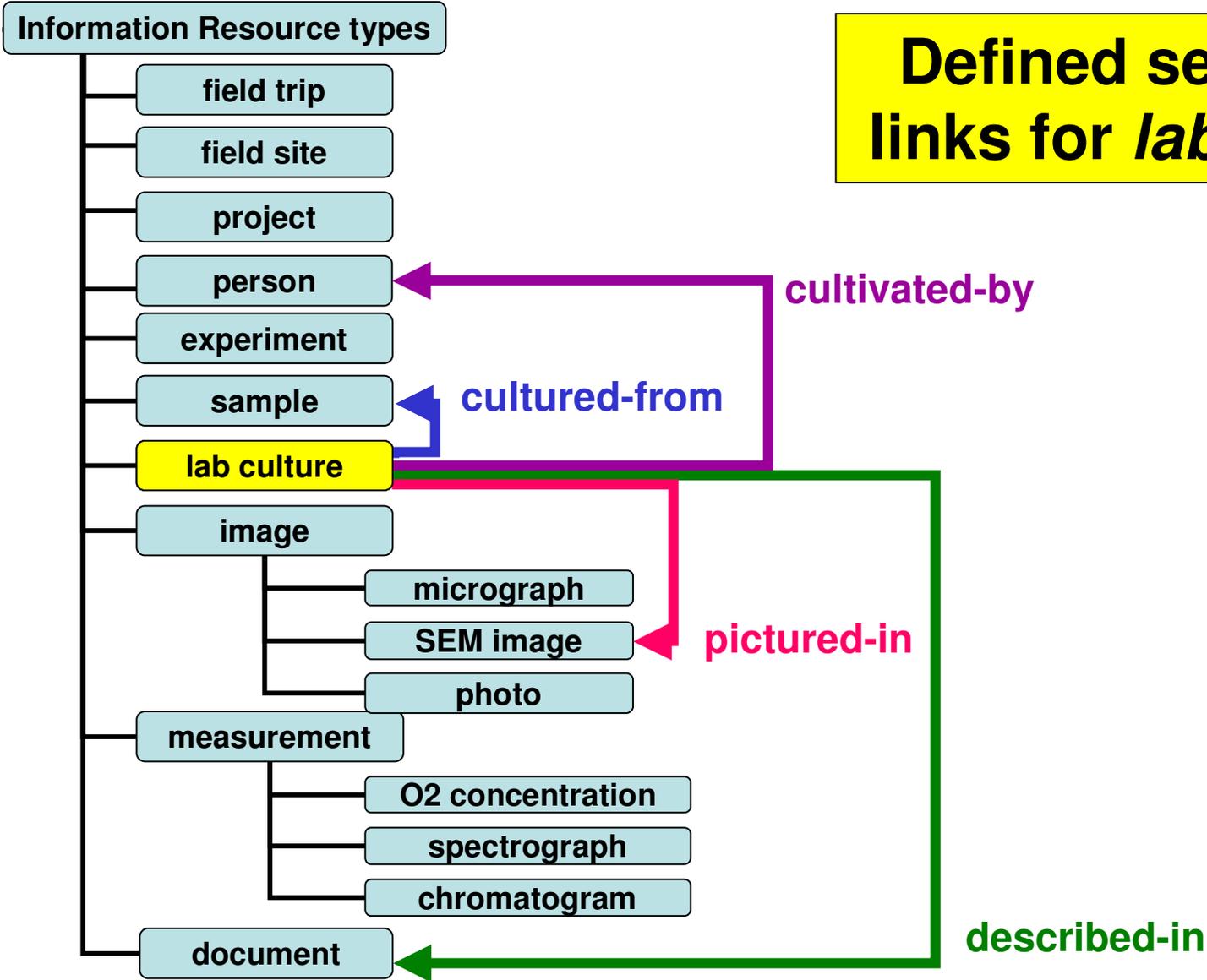
Network of Project Information



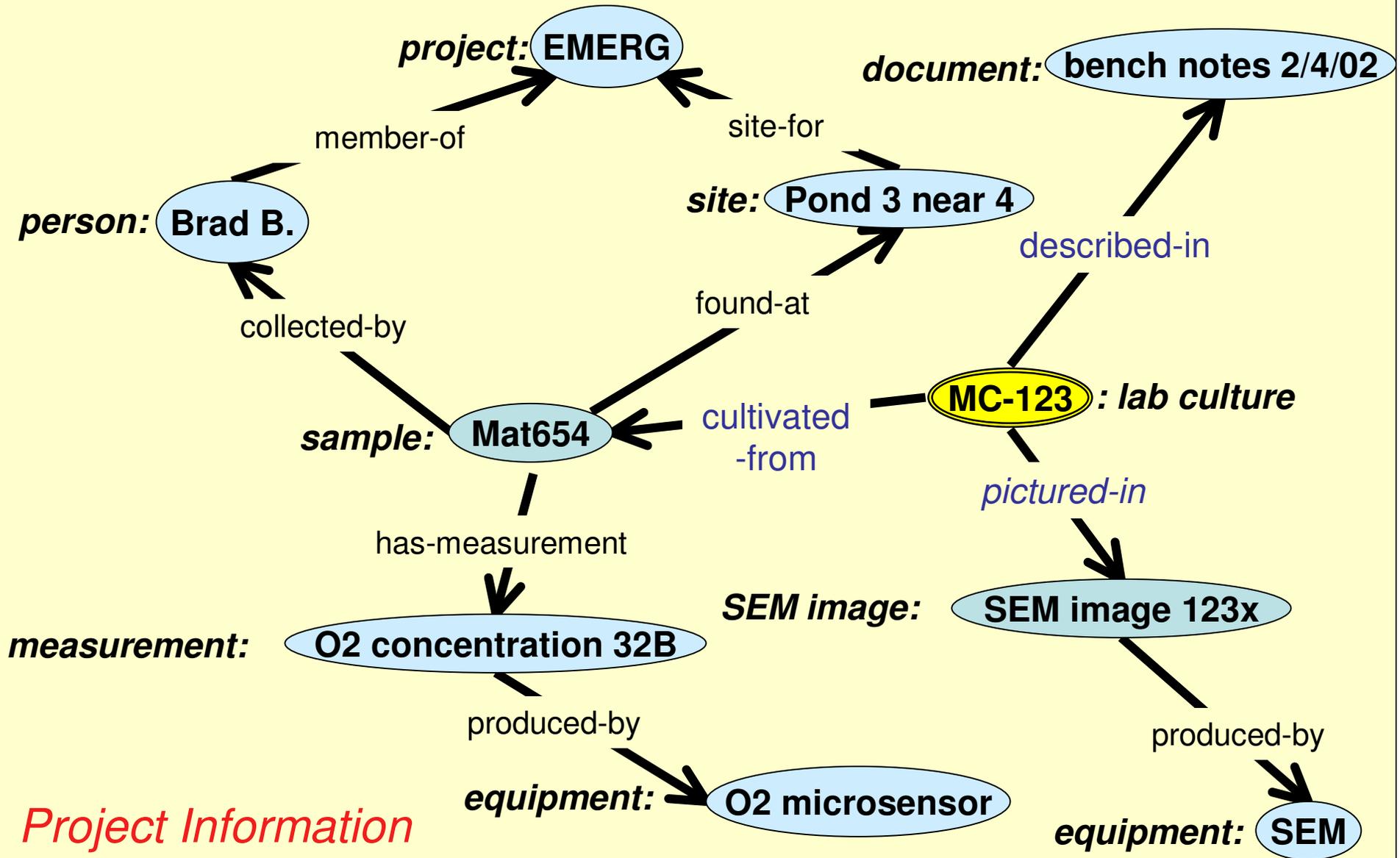
- **Nodes**: key project info resources or organizational structures (describe people, places, measurements, instruments)
- **Attributes**: properties of resources
 - **Links**: relationships among resources
- **Attached files**: electronic products associated with resources

Ontology of Project Information Resources

Defined semantic links for *lab culture*



Evolving Web of Project Information



*Project Information
 in context*

search for records *create new links* *icon identifies record type* *modify records*

create new records *Web-based, platform independent access*

Links to Related Records

- convenient navigation
- predefined links
- information traceback

click to navigate

The screenshot shows a web browser window with the ScienceOrganizer interface. The main content area is divided into two panes. The left pane, titled 'View Links for Current Item', shows a tree structure of related records including 'Stromatolite Beach: cultures', 'Cultivated From', 'Has Genetic Sequence Info', 'Has Growth Medium Recipe', 'Has Maintenance Medium Recipe', 'Isolated By', and 'Pictured In'. The right pane, titled 'Info for Current Item', shows details for 'Culture: HBC-2', including a description, cultivation data, and physiological characteristics. Navigation buttons like 'Modify', 'Enclose', 'Delete', and 'Duplicate' are visible. A 'data fields' label points to the right pane's content.

Project Information Record

- images
- datasets
- *cultures*
- samples
- field sites
- measurements
- instruments
- lab notes
- publications
- spreadsheets

Enables creation and linking of nodes, uploading of electronic files



Outline



- **ScienceOrganizer: brief introduction**
- **FY02 Work:**
 - **Integration with Work Tools & Systems**
 - **Automated Knowledge Acquisition**
 - **New Customers / scale-up**



Integration w/Work Tools & Systems



Goal: Increase usage & information capture

Strategy:

- Reduce usage barriers
- Piggyback off of existing tools and systems
- Provide unique value-added services to encourage usage
- Support current work practice, don't fight it!

ScienceOrganizer integration targets:

1. **Organizer Mail:** capture project mail & attachments
2. **MS Office Interoperation:** simplify document updates
3. **Real-time Experimentation:** enable laboratory automation
& results capture



1. Organizer Mail



- **Enable email distribution service within Organizer (a list server)**
 - **Create new discussion lists**
(list@sciencedesk.arc.nasa.gov)
 - **Manage list membership**
 - **Distribute mail to subscribers**
- **Create message archive within Organizer**
 - **Each message is a node within Organizer network**
 - **Can link email messages to items within Organizer**



Linked Email



ScienceOrganizer: An Information-Sharing Tool For Scientific Project Teams NASA Ames/Computational Sciences

[New Item](#) [Find Items](#) [Home](#) [Go To](#) [Logout](#) [Help](#)

[View Links](#) [Edit Links](#) [Modify Item](#) [Modify Permissions](#) [Delete Item](#) [Duplicate Item](#) [Put Item In Folder](#)

March Baja trip logistics (open all | close)

- Followed By (1 Email Messages)
 - Greenhouse Mat Experiment and T
- Preceded By (1 Email Messages)
 - Re: Baja, again
- Sent By (1 Persons)
 - Des Marais, David
- Sent To (1 Persons/Mailing Lists)
 - EMERG General Mailing List
- Other Permissible Links...
- Contained By (0 Email Message Fo

Email Message: March Baja trip logistics

Item ID# 60371 updated 8/1/02@11:14AM
Send this Item's web address via [Email](#)

'From:' Line David Des Marais <d-desmarais@mail.arc.nasa.gov>
Sender Des Marais, David
'To:' Line emerg@sciencedesk.arc.nasa.gov, "David A. Stahl" <d-stahl@nwu.edu>, "John R. Spear" <John.Spear@Colorado.EDU>, SOGIN@evol5.mbl.edu
Recipients EMERG General Mailing List
Date sent
Date received 2001-02-16 00:00:00
Body

Dear Group,

We are still looking forward to a March 5 to 18 schedule for the Baja trip. Last week, our Mexican research permit was recommended forward from SEMARNAP, Mexico's Department of the Environment and Fisheries, to SRe, their State Department. This week, the State Department reviewed the package, received inputs from other relevant agencies, and made additional requests from me with which I have now complied. I expect that we will receive final approval within days.

Accordingly, if you feel that making plane reservations now will save you more funds than you might lose with a postponement, please go ahead and make those reservations.

TRAVELING TO MEXICO

If you are taking a plane flight to join the trip, you are a member of the "Airplane Group." You should plan to meet in San Diego on March 5 and ride down together in a passenger van rented from Pearson Ford in San Diego. It might not be necessary for the Airplane Group to rendezvous with the Ames Group, unless some of you want to transfer equipment to our vans for customs declaration. We currently are holding a 12 passenger van in my name, but I must transfer that reservation to someone in the "Airplane" group as soon

Links

Next step:
Content-based links

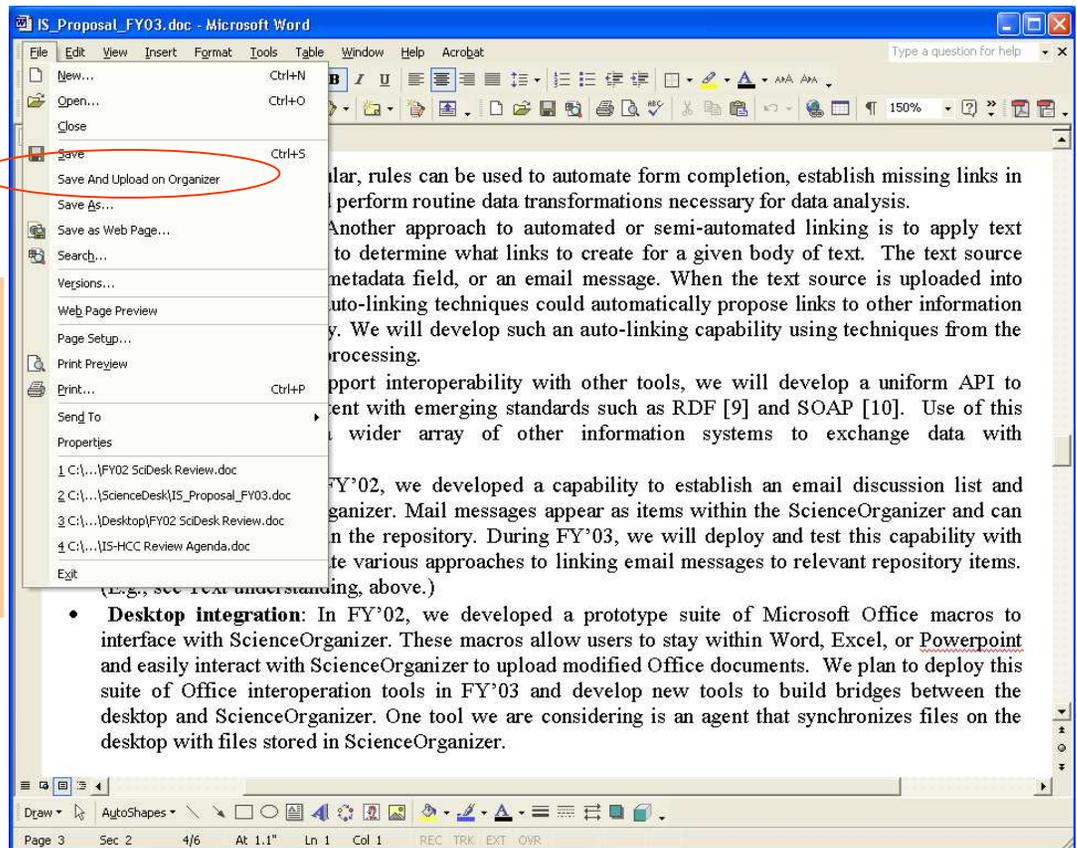
Email Body

2. MS Office Interoperation

- Developed suite of Microsoft Office macros to enable direct upload and subsequent modification of documents without leaving Office application (Deepak Kulkarni)

“Save and upload to Organizer”

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



- **Desktop integration:** In FY'02, we developed a prototype suite of Microsoft Office macros to interface with ScienceOrganizer. These macros allow users to stay within Word, Excel, or Powerpoint and easily interact with ScienceOrganizer to upload modified Office documents. We plan to deploy this suite of Office interoperation tools in FY'03 and develop new tools to build bridges between the desktop and ScienceOrganizer. One tool we are considering is an agent that synchronizes files on the desktop with files stored in ScienceOrganizer.



3. Real-time experimentation



- **Goals:**
 - Encourage use of Organizer in conjunction with day-to-day labwork
 - Enable remote experimentation in the Greenhouse laboratory (collaboratory!)
 - Facilitate sharing of laboratory results
 - Explore use of agents as assistants in the lab

Astrobiology Greenhouse Facility

Greenhouse Offers:

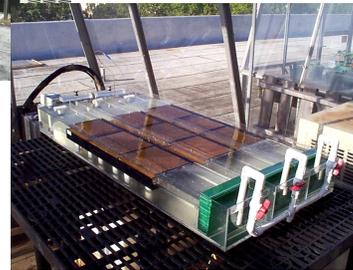
- Natural Light
- Realistic Flow Fields
- Natural Solar UV



Transport to lab



Maintenance of Field Collected Mats and Growth of Defined Communities



Experimental Manipulations

Trace Gas Production & Consumption Under "Early Earth" Conditions



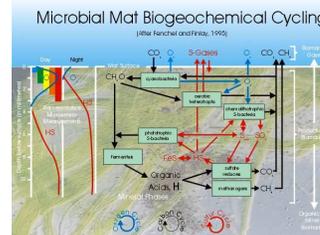
In-field collection of microbial mats



A microbial mat (+ finger)

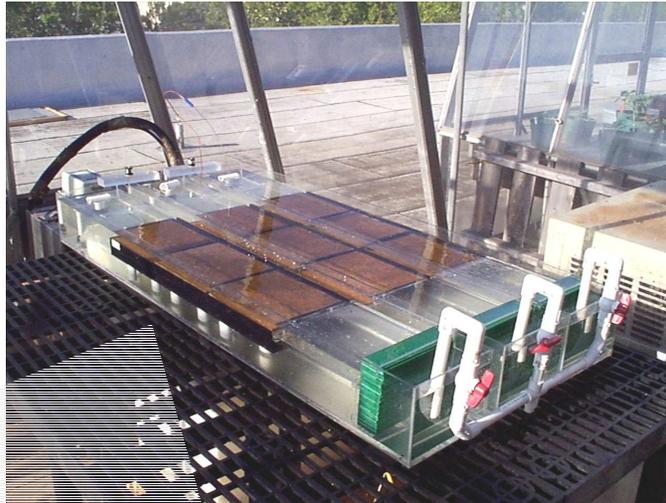
Long Term Monitoring

Detailed Studies of Mat biogeochemistry with Improved Access to Analytical Instrumentation



Experimentation in the Greenhouse *(circa 1999)*

Flume System
 with mats

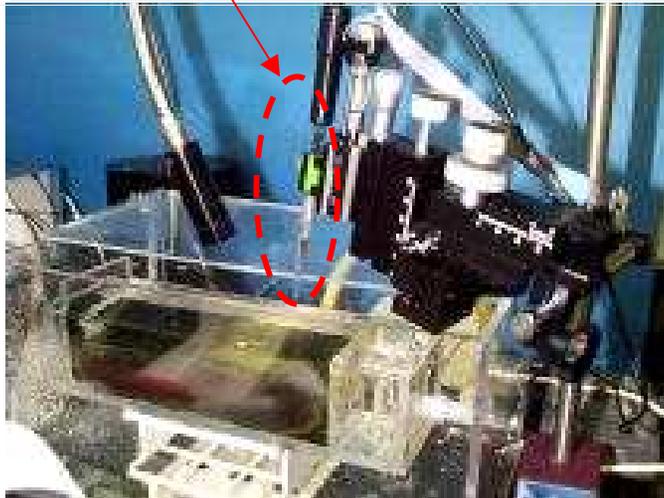


Limitations:

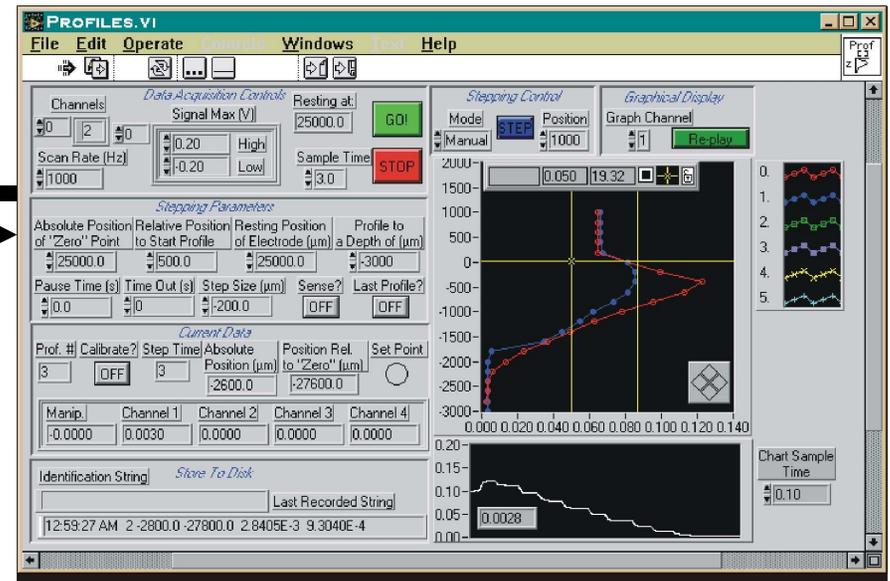
- 24x7 experiments
- remote collaborators
- physical setup

Microsensor

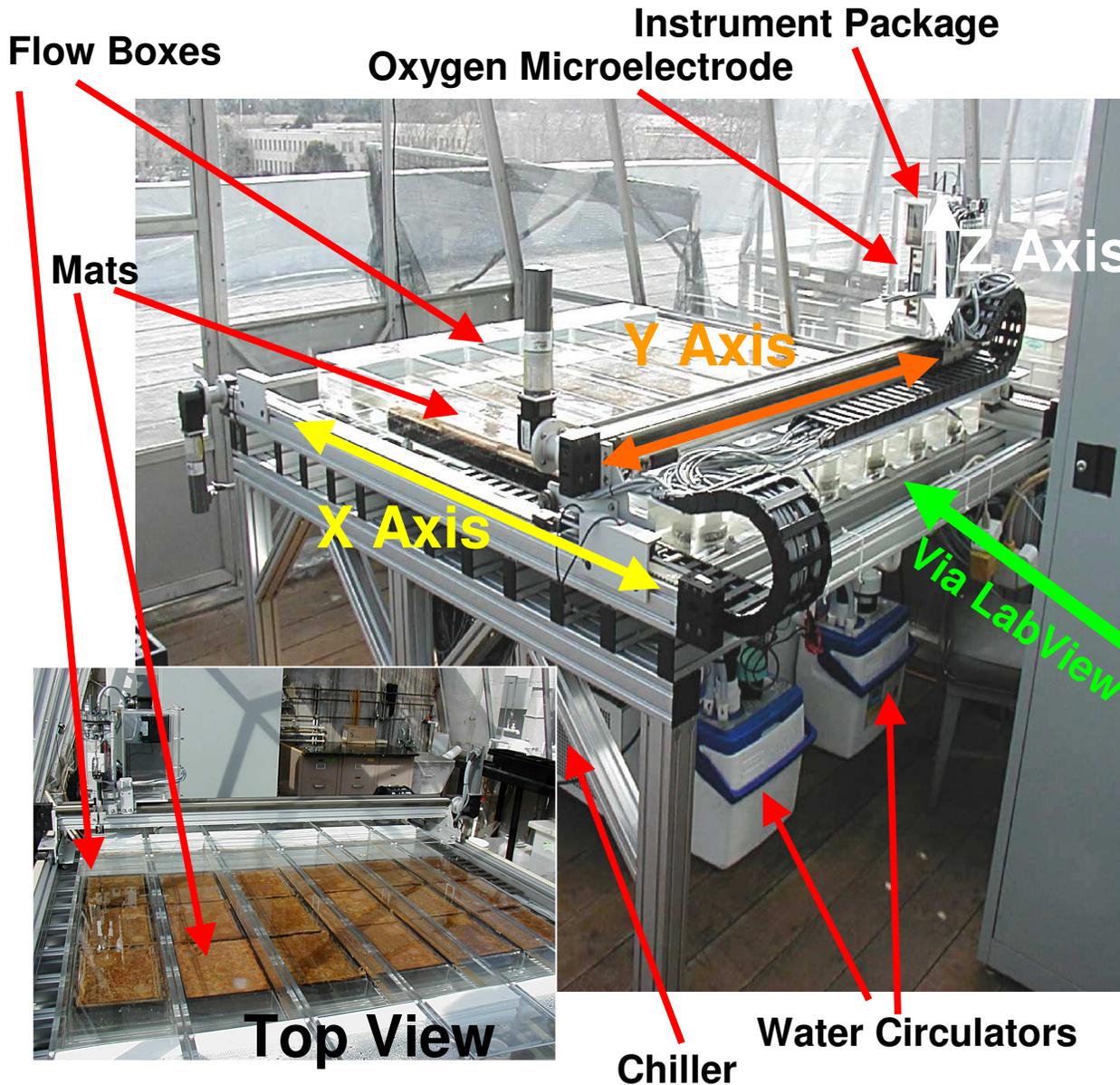
LabView Code to control
 microsensor and acquire data



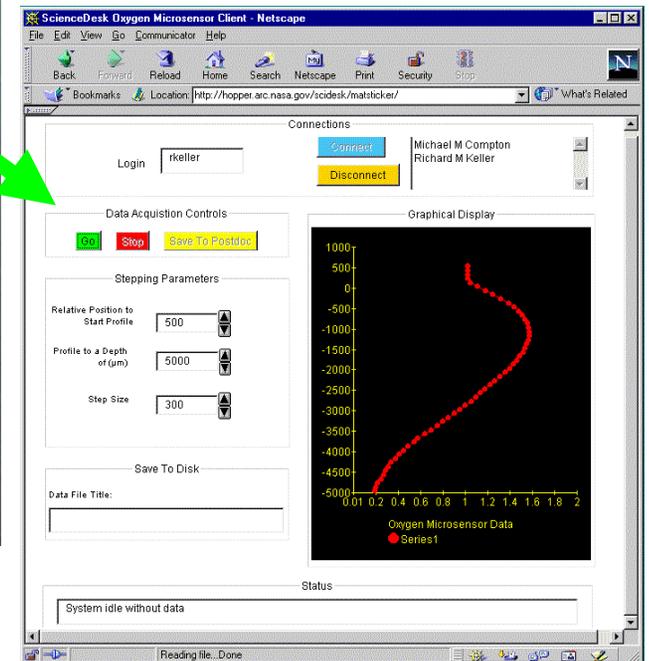
Microsensor to measure
 O_2 concentration in mat



Experimentation in the Greenhouse (circa 2001)



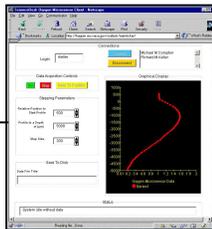
via Web/
Java applet



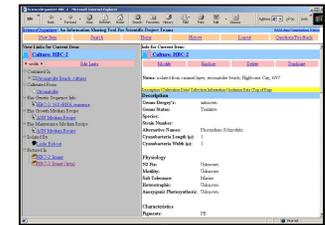
Limitations

- **Manual initiation of measurements**
- **Only a single measurement per interaction**
- **Needed:**
 - **Automated control of measurements**
 - **Multiple measurements per pass;
multiple passes repeated over time**
 - **Images too**

Experimentation in the Greenhouse (circa 2002)

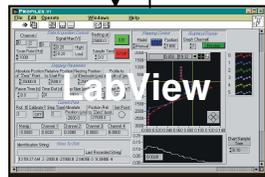


Experiment setup & activation **1**



ScienceDesk Server

7 results
6 execute



5 command
8 pass back results to repository

Table Controller Agent

4 request measurement or image

Greenhouse Measurement Pass Agent

3 spawn & initiate

Greenhouse Imaging Pass Agent

Experiment Monitoring/Scheduling Agent

2 poll

Science Organizer Repository

First user-initiated experiment successfully completed last week!

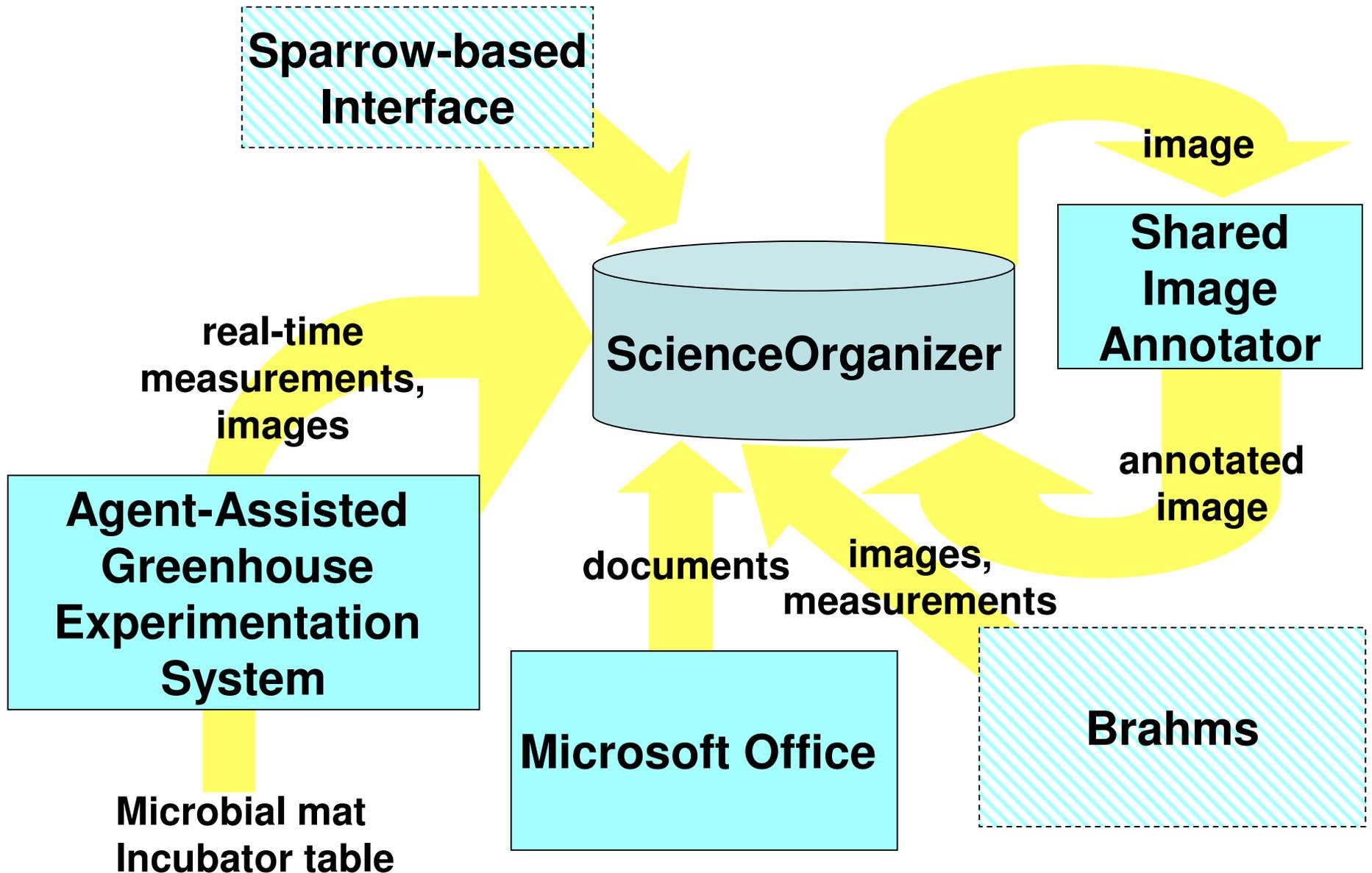


FY03 Focus Areas: Integration



- **Develop unified Organizer API to be used by all current (& future) clients**
 - **Provide essential Organizer storage and retrieval functionality**
 - **XML/RDF-based**
- **Generalize Greenhouse agent architecture**
- **Develop Windows file system mirroring functionality**

ScienceOrganizer clients





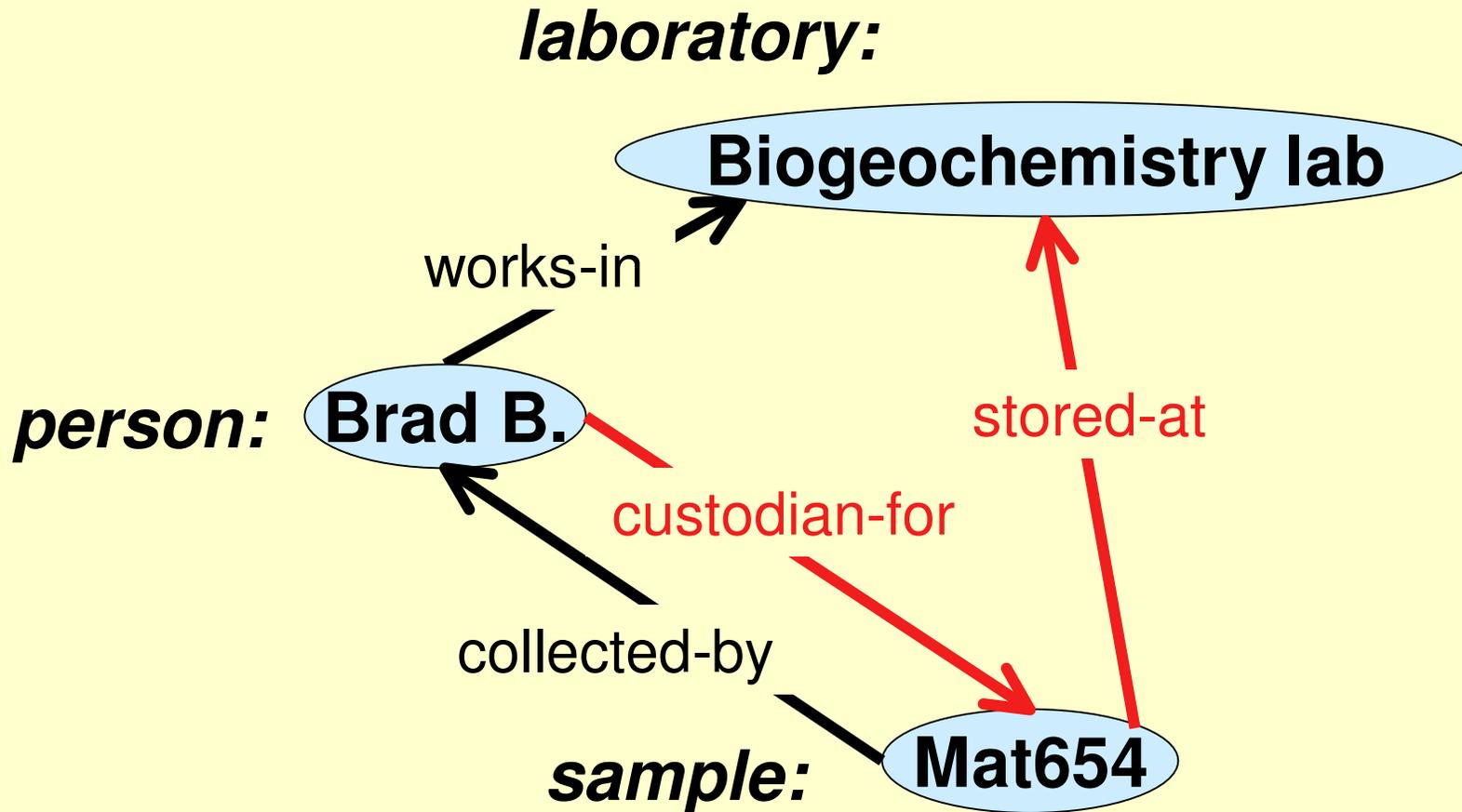
Outline



- **ScienceOrganizer brief introduction**
- **FY02 Work:**
 - **Integration with Work Tools & Systems**
 - **Automated Knowledge Acquisition**
 - **New Customers / scale-up**

- **Automated acquisition of:**
 - **Nodes**
 - **Attribute values**
 - **Links**
 1. *Via rule-based inference*
 2. *Via text analysis*

1. Rule-based Linking

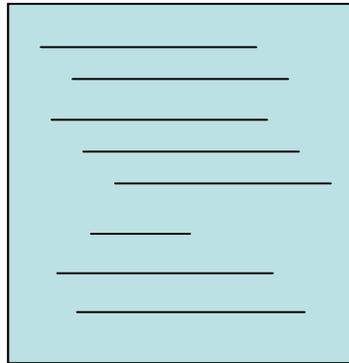


If person P is *custodian-for* sample S, and P *works-in* laboratory L,
Then S is *stored-at* L

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

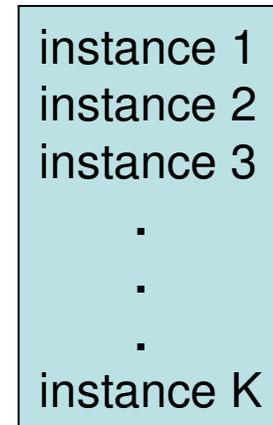
2. Text-based Linking

Given:



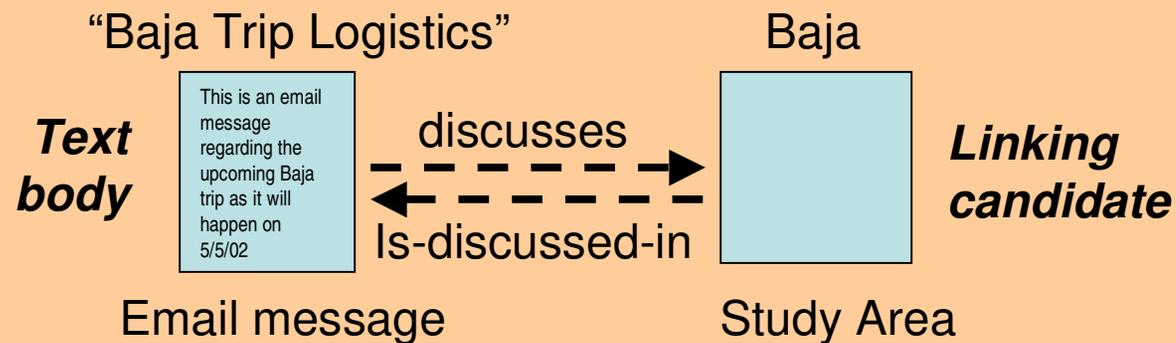
A body of text
(e.g., an email message body or attachment, text field, or document)

Find:



Linking candidates: An ordered ranking of Organizer node instances most relevant to the text

Example:





Text Analysis Approach



- **Preliminary approach by Dan Berrios based on text analysis using NLP and IR techniques**
- **Utilizes:**
 - **lexical matching**
 - **knowledge from Organizer ontology**
 - **linguistic & semantic knowledge from WordNet**



Detecting Linking Candidates through References to Names and Attribute values



- **Candidates may be identified in text by name:**
 - “Brad Bebout”
 - “Streaming Mat I”
- **...or by referencing values of their attributes:**
 - “rkeller@mail.arc.nasa.gov”
 - “604-3388”
- **and be given added credence with appearance of nouns and verbs that are appropriate to the semantics defined for the node type**
 - “Brad will prepare the sample.”
 - “I received the message from Keller.”

Example: Linking email text body

Email body (extract)

"Straw **sampling** Protocols

In situ **measurements** of fluxes:

- ... **number** of diel **periods** : 2 (once in Pond 4 near 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** : 50 mL
- ... Gas **volume taken** per **sampling interval** 5 mL"

Text marked up to identify:

- near-exact instance matches
- attribute value matches
- node/link type matches
- WordNet synonym/hyponym matches

rank	instance	class
1.063660	Spear, John	(participan)
1.063660	Stahl, David	(participan)
1.063660	Bebout, Brad	(participan)
1.036137	Ames	(study_area)
1.036137	Baja	(study_area)
1.031293	Ecogenomics Focus	(project)
1.020290	Pond 5 near 6	(aqueous_si)
1.020290	Pond 4 near 5	(aqueous_si)
1.020290	Lagoon	(aqueous_si)
1.009677	Des Marais, David	(participan)
1.009677	Farmer, Jack	(participan)
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)
1.000006	d	(document)
1.000000	major ions	(class)
1.000000	volatile sulfur	(class)
1.000000	Keane Mine Spring	(volatile_s)
0.732436	sample test	(sample)
0.669582	aqueous site	(class)
0.669582	geological site	(class)
0.666667	greenhouse experim	(class)
0.666667	SB-001	(stromatoli)



FY03 Focus Areas: Automated Knowledge Acquisition



- **Enhanced knowledge acquisition capabilities**
 - **Implement and evaluate quality of email auto-linking**
 - **Auto-attachment insertion**
 - **Auto-attribute setting**



Outline



- **ScienceOrganizer: brief introduction**
- **FY02 Work:**
 - **Integration with Work Tools & Systems**
 - **Automated Knowledge Acquisition**
 - **New Customers / scale-up**



New Customers



- **Growth in customer base**
- **Ontology editor**
- **Spin-off: InvestigationOrganizer**



Growth in user base during FY02



ScienceOrganizer

FY01

- **ARC Microbial Ecosystems Group (Code S):** field & lab science, experiments, data analysis.
- **NAI Ecogenomics Focus Group (Code S):** cross-discipline collaboration, data analysis.
- **ARC Electron Microscopy Lab (Code S):** electron microscopy image archiving, sample cataloging.

-
- **JSC Astrobiology Institute for the Study of Biomarkers (Code S):** electron microscopy image archive, sample collection, cataloging, and storage; support for education & outreach.
 - **NIH/NASA Malaria Control Study (via Fundamental Biology Program – Code S):** African malaria study - data collection and archiving.
 - **ASU/NSF Desert Microbial Survey (NSF):** microbial survey; provides publicly-accessible repository.

FY02

InvestigationOrganizer

- **CONTOUR Mishap Investigation Board (Code Q):** mishap investigation, evidence collection and archiving, failure analysis.
- **Airshow 2002 Mishap Investigation Board (Code Q):** mishap investigation, evidence collection and archiving, failure analysis.



Organizer ontology editor



- **To support rapid ontology expansion, developed an ontology editor**
 - **Allows knowledge modelers to manipulate ontology without assistance**
 - **Uses Organizer interface to edit ontology**
 - **Current version is somewhat primitive**
- **Have investigated use of existing ontology editors, e.g. Protégé**

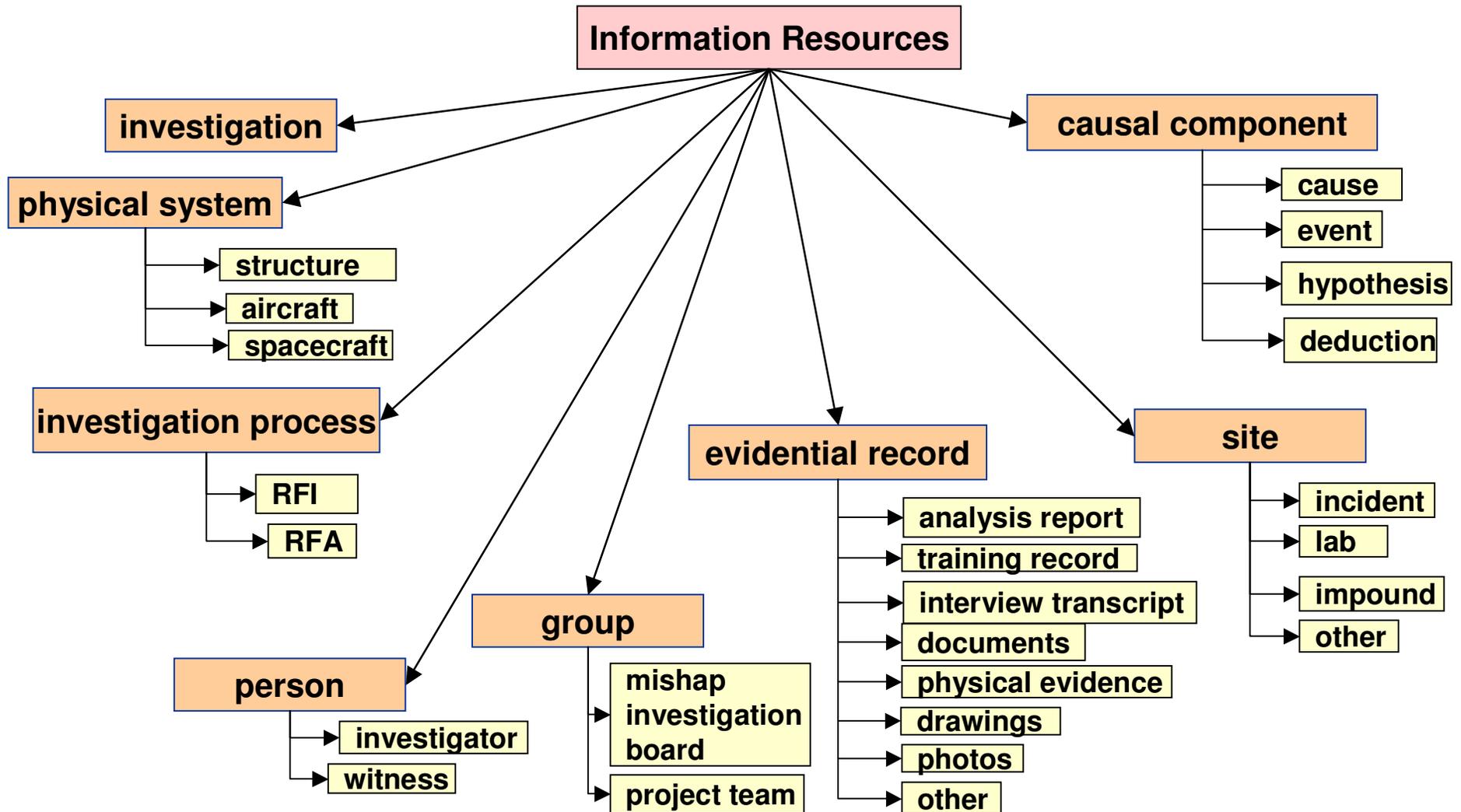


InvestigationOrganizer

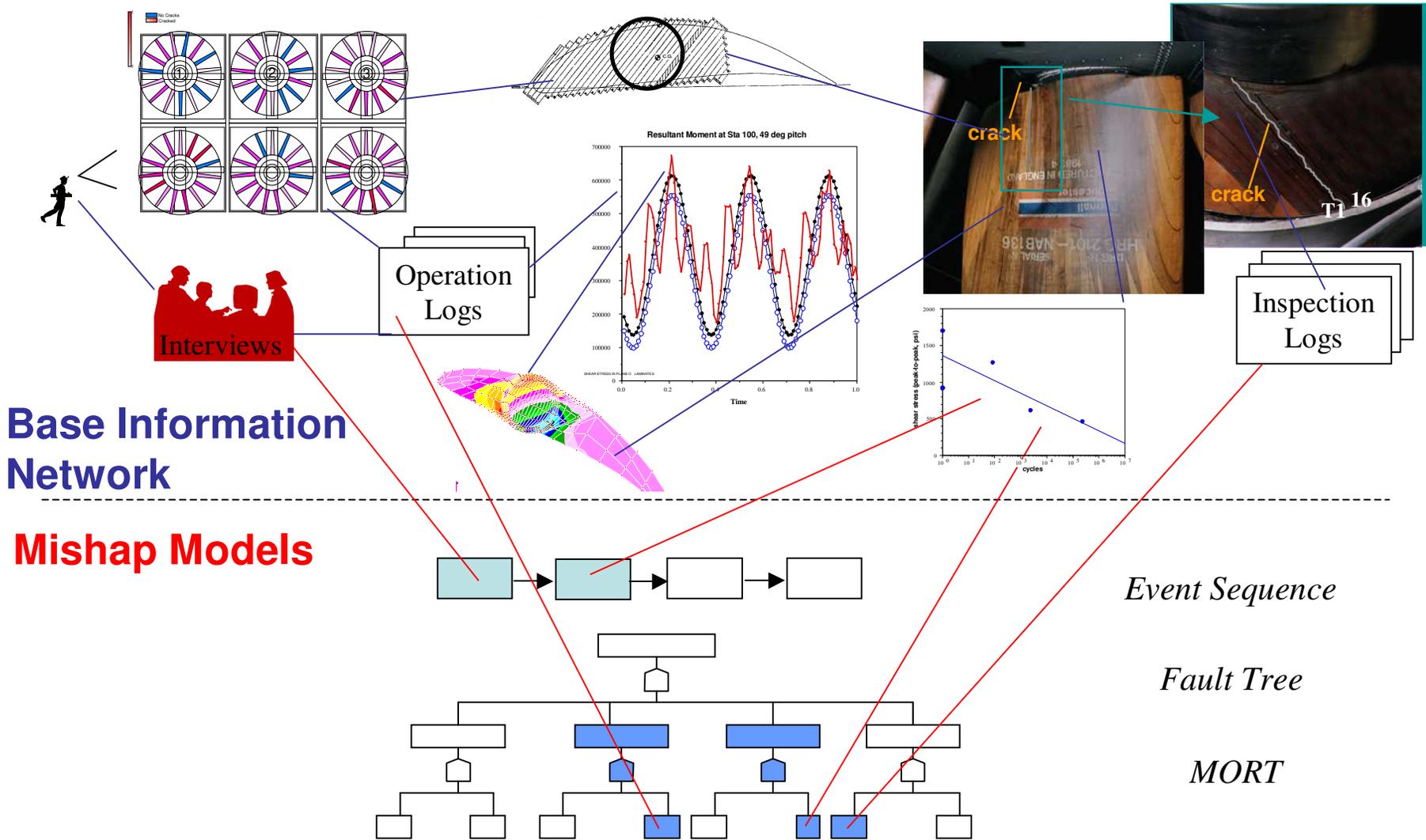


- **A customization of the ScienceOrganizer system to NASA Mishap Investigations. Same base semantically-structured information repository functionality, but:**
 - application to new domain (engineering investigations) with development of new ontology terms/relations
 - addition of a causal modeling infrastructure for mishap analysis: MORT tree, event sequences, fault tree
 - addition of visualization tools for causal model display
- **Co-funded by the Engineering for Complex Systems Program (Tina Panontin, Co-I)**

Investigation Ontology



Superimposed Mishap Models





Recent Investigation Organizer Developments



- **Pilot usage with actual investigations:**
 - **Moffett Airshow (Class C):** Damage and injury resulting from pyrotechnic display at airshow
 - **CONTOUR Mission (Class A):** Loss of CONTOUR spacecraft after solid rocket motor burn



FY03 Focus Areas: Scaling up



- **Workspaces: partitioning/filtering the info space (for different task contexts)**
- **Visualizing the info space**
- **Customizing the unified ontology to conform to project-specific terminology and information requirements**



Metrics



- **Quantitative/objective**
 - **Usage statistics**
 - Over 150 registered users
 - Over 4000 information nodes
 - Over 1500 documents stored
- **Qualitative/subjective**
 - **User evaluation & interview**
 - CONTOUR user debrief study



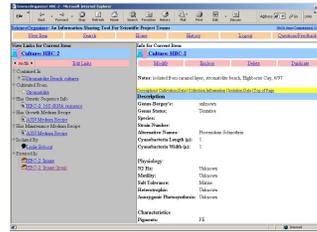
Extra Slides



ScienceOrganizer Components

Interface Layer

Node Browser Interface



Programmatic/Agent Interface



Representation & Reasoning Layer

Request Manager

Information Ontology
(node types, link types, attributes, rules)

Semantic Network
(nodes, links, attribute values)

Inference Engine

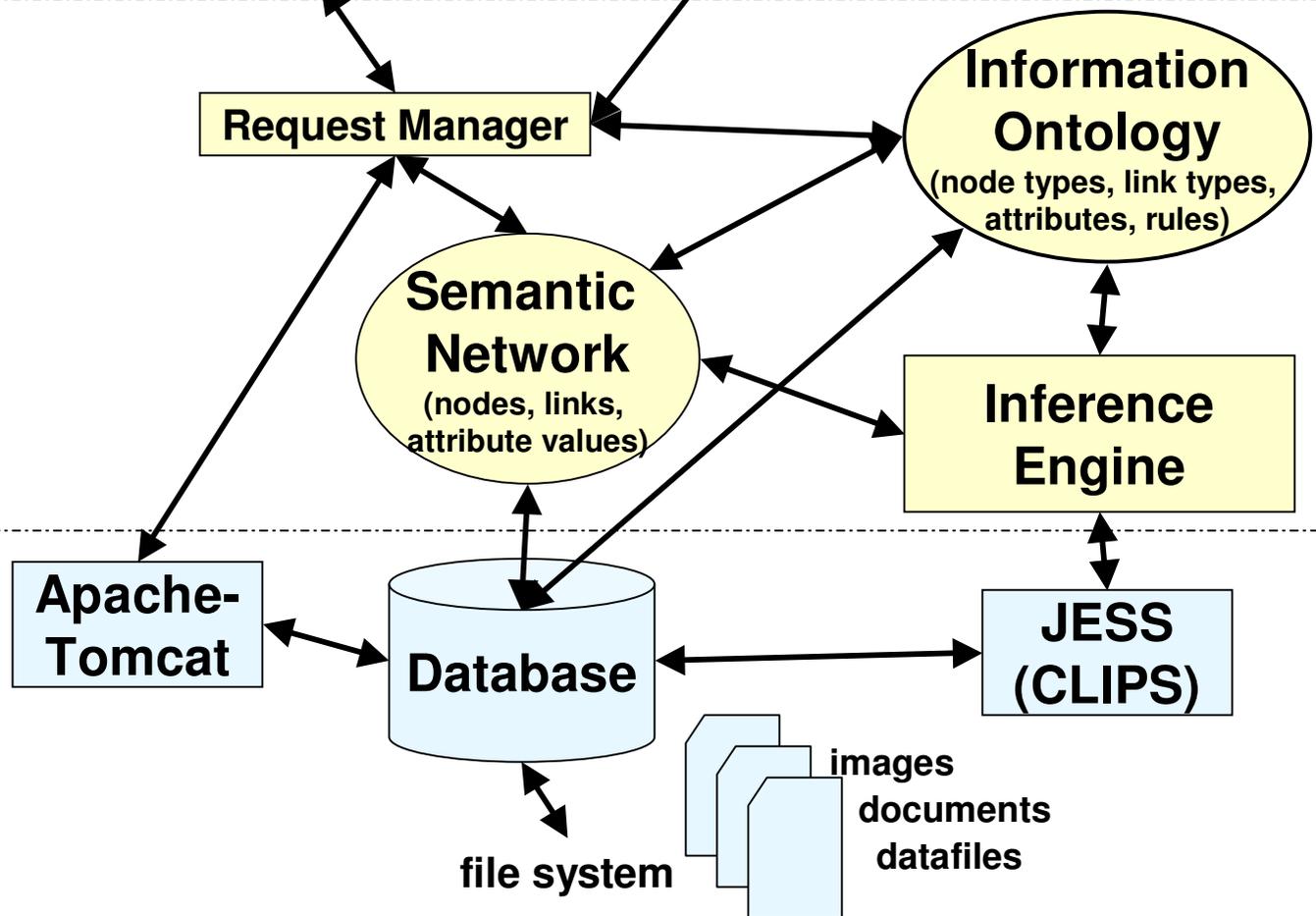
Implementation Layer

Apache-Tomcat

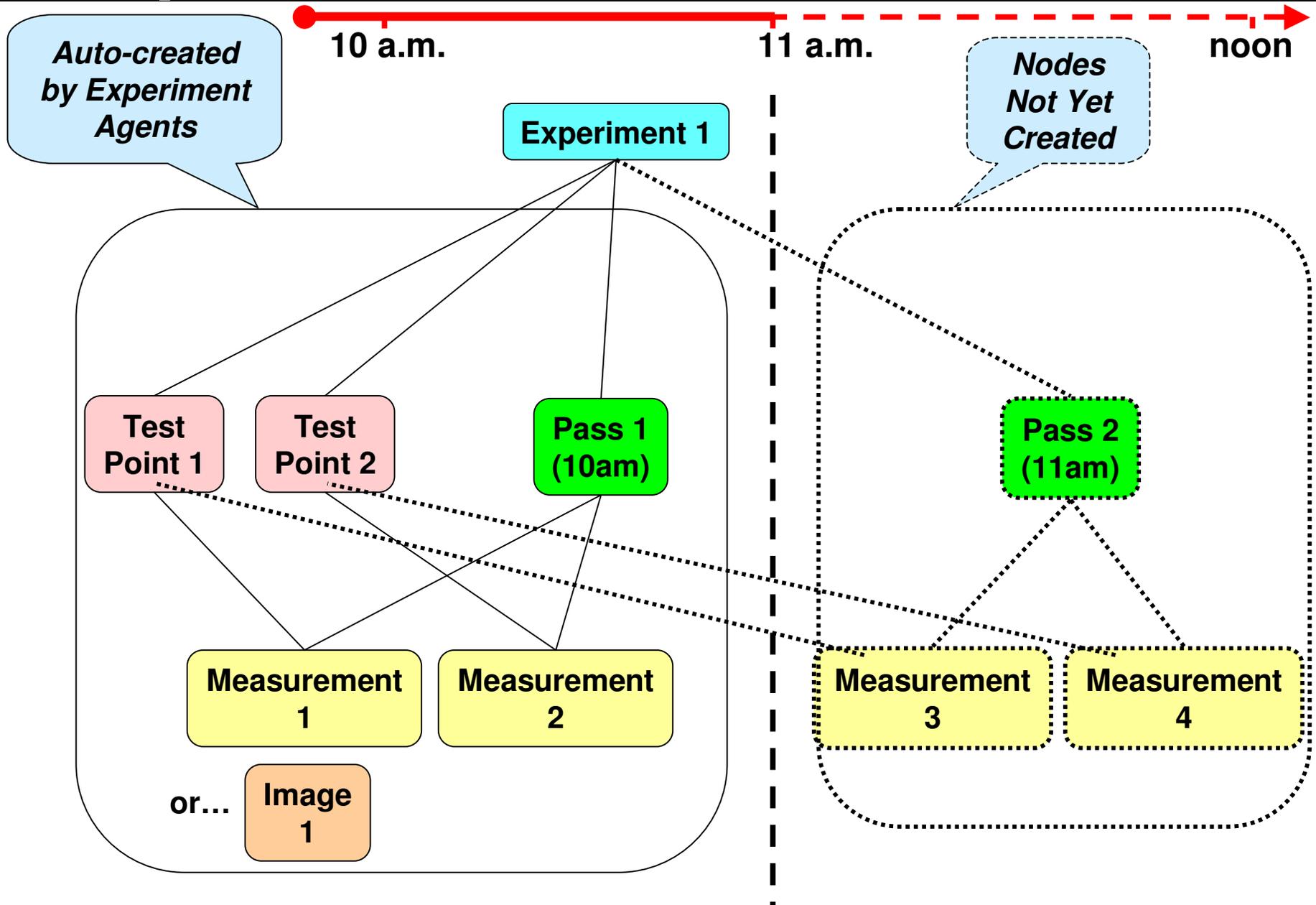
Database

JESS (CLIPS)

file system



Experiment Structure in Organizer





Link Candidate Generation Algorithm



1. **One time only** (*not automated*):

- **Map Organizer node & link types to WordNet synsets**

2. **For each text body:**

- **Find all instance names & attribute values in text body**
- **Score each candidate instance based on frequency of name/value occurrences**
- **Parse text into noun and verb phrases**
- **Match phrases to Organizer node & link types using WordNet network**
- **Increase candidate scores based on node/link type matches**
- **Propagate scores to neighboring instances in the network**

Investigation network structure with superimposed causal structure

