Making Metadata More Comprehensive and More Searchable with CEDAR

Mark A. Musen, M.D., Ph.D.
Stanford University
musen@Stanford.EDU
Welcome to our 48th Year of Publication
An (in)famous commentary in *Nature* (Begly and Ellis, 2012)

- Amgen could reproduce the findings in only 6 of 53 “landmark” papers in cancer biology
- Bayer could validate only 25% of 67 pre-clinical studies
- Some “landmark” studies had spawned entire fields of investigation, with no one bothering to confirm the initial results
- Nonreproducible studies are more likely to be published in journals with high impact factors
The problem has many causes

- A lack of statistical power in many studies
- An “art form” in conducting experiments
- An eagerness to publish early
- Outright fraud
- A system that does not make it easy or rewarding to replicate the results of other investigators
At a minimum, science needs

• Open, online access to experimental data sets
• Annotation of online data sets with adequate metadata (data about the data)
• Use of controlled terms in metadata annotations
• Mechanisms to search for metadata to find relevant experimental results
To make sense of the metadata ...

• We need standards
  – For the kinds of things we want to say about the data
  – For the language that we use to say those things
  – For how we structure the metadata

• We need a scientific culture that values making experimental data accessible to others
How can we make sense of microarray data?

- What was the substrate of the experiment?
- What array platform was used?
- What were the experimental conditions?

DNA Microarray
**Minimum Information About a Microarray Experiment - MIAME**

**MIAME** describes the **Minimum Information About a Microarray Experiment** that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. [Brazma et al., Nature Genetics]

The six most critical elements contributing towards MIAME are:

1. The raw data for each hybridisation (e.g., CEL or GPR files)
2. The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
3. The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
5. Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
6. The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data)

For more details, see **MIAME 2.0**.

MIAME does not specify a particular format, however, obviously the data are more usable, if it is encoded in a way that the essential information specified by MIAME can be accessed easily. FGED recommends the use of **MAGE-TAB** format, which is based on spreadsheets, or **MAGE-ML**.

MIAME also does not specify any particular terminology, however for automated data exchange the use of standard controlled vocabularies and ontologies are desirable. FGED recommends the use of **MGED Ontology** for the description of the key experimental concepts, and where possible ontologies developed by the respective community for describing terms such as anatomy, disease, chemical compounds etc (see **OBO page** for more detail).
• **MIAME** structures metadata for online repositories such as the Gene Expression Omnibus (GEO) and Array Express

• Provides a framework for assuring that scientists can understand the basics of what experiment was performed

• Offers a model that lots of groups have copied
A curated, informative and educational resource on inter-related data standards, databases, and policies in the life, environmental and biomedical sciences.
Minimum Information About a Microarray Experiment

Abbreviation: MIAME

@mbi REPORTING GUIDELINE

General Information

MIAME is intended to specify all the information necessary for an unambiguous interpretation of a microarray experiment, and potentially to reproduce it. MIAME defines the content but not the format for this information.

Homepage http://www.fged.org/projects/miame/
Developed in United Kingdom, France, Germany, Netherlands, Belgium, United States of America
Created in 1999
Taxonomic range

Scope and data types

- Microarray Data
- Genome
- DNA
- DNA Microarray
- Transcriptome
- RNA
- Nucleic Acid Hybridization

Record updated: March 11, 2016, 5:33 p.m. by The BioSharing Team.

Recommended by

- EMBO Press
- Scientific Data

In Collections

- National Child Development Study (UK)
- DNA Microarray

Related Standards

Reporting Guidelines

- Minimum Information about an ENVironmental transcriptomic experiment
- Minimal Information about a high throughput SEQuencing Experiment
- Minimum Information About a Microarray Experiment involving Plants
- Minimum Information about a Nutrigenom experiment
- Minimum Information about a array-based

Implementing Databases (4)

ArrayExpress
ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MiNSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

Gene Expression Omnibus
The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the studies and gene expression patterns stored in GEO.

The Immunome Database and Analysis Portal

Implementing Policies

- Scientific Data's Recommended Data Repositories
- EMBO Press Recommended Databases and Data Standards
BioSharing lists more than 100 guidelines like MIAME for reporting experimental data.
The Good News: Minimal information checklists, such as MIAME, are being advanced from all sectors of the biomedical community.

The Bad News: Investigators view requests for even “minimal” information as burdensome; there is still the pervasive problem of “What’s in it for me?”
Submitting to GEO

Summary Data Matrix

Raw Data

Submission

This image cannot currently be displayed.
# Use this template for 3' or whole Gene expression studies when summarization probe set data will be provided as CHP files.
# Do NOT submit CHP files unless they are relevant to your analysis (instead, use the Matrix table option to submit the relevant data, e.g. Bioconductor).
# Incomplete submissions will be returned. Click the Metadata Example tab below to view a completed worksheet.
# A complete submission will consist of: (1) a completed metadata worksheet, (2) the CHP files, and (3) the original CEL files.
# Field names (in blue on this page) should not be edited. Hover over cells containing field names to view field content guidelines or,
# CLICK HERE for Field Content Guidelines Web page.

**SERIES**
# This section describes the overall study.

<table>
<thead>
<tr>
<th>title</th>
<th>summary</th>
<th>overall design</th>
<th>contributor</th>
<th>contributor</th>
</tr>
</thead>
</table>

**SAMPLES**
# The Sample names in the first column are arbitrary but they must match the column headers of the Matrix table (see next worksheet).

<table>
<thead>
<tr>
<th>Sample name</th>
<th>title</th>
<th>CHP file</th>
<th>source name</th>
<th>organism</th>
<th>characteristics: tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMPLE 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMPLE X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PROTOCOLS**
# This section includes protocols and fields which are common to all Samples.
# Protocols which are applicable to specific Samples or specific channels should be included in additional columns of the SAMPLES section instead.

| growth protocol | treatment protocol | extract protocol | label protocol | hyb protocol |

[Optional] Describe the conditions that were used to grow or maintain organisms or cells prior to extract preparation.

Unique title (less than 120 characters) that describes the overall study.

"Firstname, Initial, Lastname". Example: "John, H, Smith" or "Jane, Doe".

Unique title that describes the Sample. We suggest that you use the convention: [biomaterial]-[condition(s)]-[replicate number], e.g., Muscle_exercised_60min_rep2.

Replace 'tag' with a biosource characteristic (e.g. "gender", "strain", "tissue", "developmental stage", "tumor stage", etc), and then enter the value for each sample beneath (e.g. "female", "129SV", "brain", "embryo", etc). You may add additional characteristics columns to this template (see 'Metadata Example' spreadsheet).
Reliance on free text leads to a mess!

age
Age
AGE
`Age
age (after birth)
age (in years)
age (y)
age (year)
age (years)
Age (years)
Age (Years)
age (yr)
age (yr-old)
age (yrs)
Age (yrs)

age [y]
age [year]
age [years]
age in years
age of patient
Age of patient
age of subjects
age(years)
Age(years)
Age(yrs.)
Age, year
age, years
age, yrs
age.year
age_years
CEDAR Partners

• Stanford University School of Medicine

• University of Oxford e-Science Centre

• Yale University School of Medicine

• Northrup Grumman Corporation

• Stanford University Libraries
From standard checklists to CDE-like elements to annotation templates

- Serve machine-readable **content metadata standards**, providing **provenance for their elements**
- Inform the creation of metadata templates, rendering standards invisible to the researchers
HIPC centers submit data directly to the NIAID ImmPort data repository using detailed, experiment-specific templates.

Example entity–relationship diagram for describing metadata for annotating multiplex bead array assays (e.g., Luminex)
ImmPort’s Mission

ImmPort is a long-term, sustainable data warehouse for the purpose of promoting re-use of immunological data generated by NIAID DAIT and DMID funded investigators.

ImmPort supports analysis of flow cytometry results and HLA genetic associations.

What is ImmPort

Flow Cytometry Analysis (FLOCK)

Flow cytometry analysis component includes:
- Automated cell population identification
- Result visualization in 2D and 3D
- Statistical analysis of population characteristics
- Automated mapping of populations across multiple samples

Open ImmPort

- Browse and search for shared study data
- Cytokine and cell interaction literature mining:
  - ImmuneXpresso
- Example R and Python analysis code
- Cytokine registry
- Cell Ontology Visualizer

Data Release

March 2016 - Data Deluge - 48 Studies released! Human Immunology Project Consortium (HIPC) teams from Stanford, Baylor, Dana Farber, Yale, Emory and Seattle have contributed extensively. ImmPort is also sharing March of Dimes microbiome study data SDY465 and data from 5 RELIVE clinical trials (SDY289 to SDY292, SDY294). Combined with human T cell distribution data from the Donna Farber team of Columbia Center for Translational Immunology SDY702 and you have a MAJOR data deluge! What data can you expect to find? Multi-year flu vaccine studies, HAI, ELISPOT, CyTOF, flow cytometry, clinical assessments and much more. Get the scoop on all the new data in ImmPort Data Release notes.
A Metadata Ecosystem

- **HIPC investigators** perform experiments in human immunology
- **HIPC Standards Working Group** creates metadata templates to annotate experimental data in a uniform manner
- **ImmPort** stores HIPC data (and metadata) in its public repository
- **CEDAR** will ease
  - Template creation and management
  - The use of templates to author metadata for ImmPort
  - Analysis of existing metadata to inform the authoring of new metadata
The CEDAR Approach to Metadata
CEDAR technology will give us

• Mechanisms
  – To author metadata template elements
  – To assemble them into composite templates that reflect community standards
  – To fill out templates to encode experimental metadata

• A repository of metadata from which we can
  – Learn metadata patterns
  – Guide predictive entry of new metadata

• Links to the National Center for Biomedical Ontology to ensure that metadata are encoded using appropriate ontology terms
The CEDAR Approach to Metadata

Authoring of Metadata Templates

- Template authors (e.g., standards committees)

  define

  Metadata templates

Annotation of Data with Metadata

  contribute

  fill in

  Metadata acquisition forms

Exploration and Reuse of Datasets through Metadata

  search, reuse

  Metadata repository

Scientists
<table>
<thead>
<tr>
<th>Title</th>
<th>Created</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIBFRAME</td>
<td>11/5/16 1:37 PM</td>
<td>11/5/16 1:37 PM</td>
</tr>
<tr>
<td>CDEs</td>
<td>8/21/16 10:22 AM</td>
<td>8/21/16 10:22 AM</td>
</tr>
<tr>
<td>Sample templates</td>
<td>11/4/16 8:19 PM</td>
<td>11/4/16 8:19 PM</td>
</tr>
<tr>
<td>BioSample Human</td>
<td>11/17/16 3:56 PM</td>
<td>11/17/16 3:56 PM</td>
</tr>
<tr>
<td>ImmPort Study</td>
<td>5/2/16 12:38 PM</td>
<td>5/2/16 12:38 PM</td>
</tr>
<tr>
<td>Investigation</td>
<td>5/7/16 2:03 PM</td>
<td>5/7/16 2:03 PM</td>
</tr>
<tr>
<td>LINCS Antibody</td>
<td>10/12/16 3:42 PM</td>
<td>10/12/16 3:42 PM</td>
</tr>
<tr>
<td>LINCS Cell Line</td>
<td>10/12/16 3:57 PM</td>
<td>10/12/16 3:57 PM</td>
</tr>
<tr>
<td>Title</td>
<td>Created</td>
<td>Modified</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>BIBFRAME</td>
<td>11/5/16 1:37 PM</td>
<td>11/5/16 1:37 PM</td>
</tr>
<tr>
<td>CDEs</td>
<td>8/21/16 10:22 AM</td>
<td>8/21/16 10:22 AM</td>
</tr>
<tr>
<td>Sample templates</td>
<td>11/4/16 8:19 PM</td>
<td>11/4/16 8:19 PM</td>
</tr>
<tr>
<td>BioSample Human</td>
<td>11/17/16 3:56 PM</td>
<td>11/17/16 3:56 PM</td>
</tr>
<tr>
<td>ImmPort Study</td>
<td>5/2/16 12:38 PM</td>
<td>5/2/16 12:38 PM</td>
</tr>
<tr>
<td>Investigation</td>
<td>5/7/16 2:03 PM</td>
<td>5/7/16 2:03 PM</td>
</tr>
<tr>
<td>LINCS Antibody</td>
<td>10/12/16 3:42 PM</td>
<td>10/12/16 3:42 PM</td>
</tr>
<tr>
<td>LINCS Cell Line</td>
<td>10/12/16 3:57 PM</td>
<td>10/12/16 3:57 PM</td>
</tr>
</tbody>
</table>
The CEDAR Approach to Metadata

- **Authoring of Metadata Templates**
  - Template authors (e.g., standards committees)
  - Define
  - Metadata templates

- **Annotation of Data with Metadata**
  - Contribute
  - Metadata acquisition forms

- **Exploration and Reuse of Datasets through Metadata**
  - Search, reuse
  - Metadata repository

Scientists

- Example datasets: HMP (Human Microbiome Project), IMMPORT, The Cancer Genome Atlas
Template Name: BioSample Human

Template Description: BioSample Human

- **Sample Name**
- **Organism**
- **Tissue**
  - Type of tissue the sample was taken from
<table>
<thead>
<tr>
<th>FIELD TYPE</th>
<th>VALUES</th>
<th>MULTIPLE</th>
<th>REQUIRED</th>
<th>SUGGESTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Type</td>
<td>Source</td>
<td>Identifier</td>
<td>No. Values</td>
</tr>
</tbody>
</table>

Sample Name

Organism

Type of tissue the sample was taken from

Enter Field Title

Tissue

Enter Field Description (Help Text)
Find terms in BioPortal or **Create New Terms** to constrain the values of the 'Tissue' field

500 results for the query 'Tissue'. Click on a term below to select it

<table>
<thead>
<tr>
<th>TERM</th>
<th>DEFINITION</th>
<th>TYPE</th>
<th>SOURCE</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>tissue</td>
<td>Multicellular anatomical structure that consists of many cells of one or a few types, arranged in an extracellular...</td>
<td>Class</td>
<td>UBERON</td>
<td>UBERON_0000479</td>
</tr>
<tr>
<td>tissue</td>
<td>-</td>
<td>Class</td>
<td>MA</td>
<td>MA_0003002</td>
</tr>
<tr>
<td>Tissue</td>
<td>-</td>
<td>Class</td>
<td>NIFSTD</td>
<td>binlex_19</td>
</tr>
<tr>
<td>tissue</td>
<td>Anatomical structure, that consists of similar cells and intercellular matrix, aggregated according to genetically...</td>
<td>Class</td>
<td>TAO</td>
<td>CARO_0000043</td>
</tr>
</tbody>
</table>
Multicellular anatomical structure that consists of many cells of one or a few types, arranged in an extracellular matrix such that their long-range organisation is at least partly a repetition of their short-range organisation.
The CEDAR Approach to Metadata

Authoring of Metadata Templates
Template authors (e.g., standards committees)
- define
- Metadata templates

Annotation of Data with Metadata
- contribute
- Metadata acquisition forms

Exploration and Reuse of Datasets through Metadata
- fill in
- search, reuse
- Metadata repository

Scientists

Projects:
- IMMPORT
- HMP
- The Cancer Genome Atlas
Metadata authors

fill in templates with metadata

Metadata Editor

suggest values

Value Recommender

metadata

stored

indexed

Metadata Repository

MongoDB

JSON-LD

Search Engine

Elasticsearch

real-time analysis

Search Engine

Elasticsearch

Metadata Repository

MongoDB

JSON-LD

Value Recommender

suggest values

real-time analysis

Metadata Editor

fill in templates with metadata

Metadata authors
### Sample Name
056X

### Organism
**Homo sapiens**

### Tissue
- **lung (UBERON) (5%)**
- **colon (UBERON) (5%)**
- **brain (UBERON) (4%)**
- **skin of body (UBERON) (3%)**
- **serum (UBERON) (2%)**
- **abdomen connective tissue (UBERON)**
- **abdominal fat pad (UBERON)**
- **abdominal segment connective tissue (UBERON)**

### Biomaterial Provider

### Optional Attribute

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tissue: brain
Sex: male
Isolate*: N/A
Age*: 74 years
Biomaterial Provider*: Life Technologies

Optional Attribute:
Name: disease
Value:
- Parkinson's disease (DOID) (39%)
- central nervous system lymphoma (DOID) (27%)
- autistic disorder (DOID) (22%)
- melanoma (DOID) (5%)
- Edwards syndrome (DOID) (2%)
- schizophrenia (DOID) (1%)
Tissue: lung
Sex
Isolate*
Age*
Biomaterial Provider*

Optional Attribute
Name: disease
Value:
- lung cancer (DOID) (61%)
- chronic obstructive pulmonary disease (DOID) (13%)
- lung squamous cell carcinoma (DOID) (5%)
- idiopathic pulmonary fibrosis (DOID) (4%)
- lung adenocarcinoma (DOID) (4%)
- adenocarcinoma (DOID) (3%)
- carcinoma (DOID) (2%)
- cancer (DOID) (2%)
<table>
<thead>
<tr>
<th>Sample Name*</th>
<th>056X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Tissue</td>
<td>skin of body</td>
</tr>
<tr>
<td>Sex</td>
<td>male</td>
</tr>
<tr>
<td>Isolate*</td>
<td>N/A</td>
</tr>
<tr>
<td>Age*</td>
<td>74 years</td>
</tr>
<tr>
<td>Biomaterial Provider*</td>
<td>Life Technologies</td>
</tr>
</tbody>
</table>

Optional Attribute

<table>
<thead>
<tr>
<th>Name</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>dermatitis</td>
</tr>
</tbody>
</table>
The CEDAR Approach to Metadata

Authoring of Metadata Templates
- Template authors (e.g., standards committees)
  - define
  - Metadata templates

Annotation of Data with Metadata
- Metadata acquisition forms
  - contribute
  - fill in

Exploration and Reuse of Datasets through Metadata
- Metadata repository
  - search, reuse

Scientists
Some key features of CEDAR

- All semantic components—template elements, templates, ontologies, and value sets—are managed as first-class entities.
- User interfaces and drop-down menus are not hardcoded, but are generated on the fly from CEDAR’s semantic content.
- All software components have well defined APIs, facilitating reuse of software by a variety of clients (see https://metadatacentre.org/tools-training/cedar-api).
- CEDAR generates all metadata in JSON-LD, a widely adopted Web standard that can be translated into other representations.
ImmPort's Mission

ImmPort is a long-term, sustainable data warehouse for the purpose of promoting re-use of immunological data generated by NIAID DAIT and DMID funded investigators.

ImmPort supports analysis of flow cytometry results and HLA genetic associations.

What is ImmPort

Flow Cytometry Analysis (FLOCK)
Flow cytometry analysis component includes:
- Automated cell population identification
- Result visualization in 2D and 3D
- Statistical analysis of population characteristics
- Automated mapping of populations across multiple samples

Open ImmPort
- Browse and search for shared study data
- Cytokine and cell interaction literature mining: ImmuneXpresso
- Example R and Python analysis code
- Cytokine registry
- Cell Ontology Visualizer

Data Release
March 2016 - Data Deluge - 48 Studies released! Human Immunology Project Consortium (HIPC) teams from Stanford, Baylor, Dana Farber, Yale, Emory and Seattle have contributed extensively. ImmPort is also sharing March of Dimes microbiome study data SDY465 and data from 5 RELIVE clinical trials (SDY289 to SDY292, SDY294). Combined with human T cell distribution data from the Donna Farber team of Columbia Center for Translational Immunology SDY702 and you have a MAJOR data deluge! What data can you expect to find? Multi-year flu vaccine studies, HAI, ELISPOT, CyTOF, flow cytometry, clinical assessments and much more. Get the scoop on all the new data in ImmPort Data Release 18 notes.
LINCS aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents.
Hydra-in-a-Box

One Body, Many Heads

- ETDs (Theses)
- Books, Articles
- Images
- Audio-Visual
- Research
- Maps & GIS
- Documents

Scalable, Robust, Shared Management and Preservation Services
Open data is about more than disclosure. It must be "fair".

- Findable
- Accessible
- Interoperable
- Reusable
How will CEDAR make data FAIR?

Data will be more findable because

• Datasets will be associated with rich metadata
  – Use of community-based templates will set expectations for the
    metadata elements that need to be specified
  – Use of predictive metadata entry will encourage the creation of
    more complete metadata specifications

• Metadata will use standard ontologies and value sets, that
  are themselves based on FAIR principles

• Metadata, and the data sets that they reference, will be
  given unique identifiers
How will CEDAR make data **FAIR**?

Data will be more **accessible** because

- All metadata and their associated data sets will have unique identifiers
- The CEDAR metadata archive will be persistent, even if the target data repositories are ever decommissioned
How will CEDAR make data FAIR?

Data will be more interoperable because

• CEDAR metadata are stored in a standard, widely used knowledge-representation language (JSON-LD)

• Use of community-based content standards (for templates, terminologies, and value sets) assures shared representation of metadata
How will CEDAR make data FAIR?

Data will be more **reusable** because

- Metadata will include provenance information for
  - Data sets
  - Metadata themselves—an essential component because ontologies, experimental methods, and scientific knowledge evolve

- Metadata will be more comprehensive and more complete because
  - They will be easier to author
  - They will be constructed with community buy-in, due to the use of community-developed standards
Where is this ecosystem leading?

• Formalization of scientific knowledge in terms of ontologies
• Use of ontologies to standardize scientific communication
• In the long term ...
  – Dissemination of the results of scientific investigations purely as machine-interpretable knowledge
  – Intelligent agents that will allow help us to synthesize scientific knowledge from online representations
Imagine a Siri-like intelligent agent on the Web ...

• That tells you when new studies have been done in disciplines that you care about
• That can compare and contrast study results across dimensions that matter to you
• That can identify those studies whose subjects best match the characteristics of particular patients
• That can tell investigators when the methods of their proposed studies are nearly identical to those of earlier studies
Once scientific knowledge is disseminated in machine-interpretable form …

• Technology such as CEDAR will assist in the automated “publication” of scientific results online

• Intelligent agents will
  – Search and “read” the “literature”
  – Integrate information
  – Track scientific advances
  – Suggest the next set of experiments to perform
  – And maybe even do them!